Multilinear Modelling of Faces and Expressions

Stella Graßhof, *Member, IEEE,* Hanno Ackermann, Sami Sebastian Brandt, *Member, IEEE,* and Jörn Ostermann, *Fellow, IEEE*

Abstract—In this work, we present a new versatile 3D multilinear statistical face model, based on a tensor factorisation of 3D face scans, that decomposes the shapes into person and expression subspaces. Investigation of the expression subspace reveals an inherent low-dimensional substructure, and further, a star-shaped structure. This is due to two novel findings: (1) increasing the strength of one emotion approximately forms a linear trajectory in the subspace. (2) All these trajectories intersect at a single point – not at the neutral expression as assumed by almost all prior works – but at an *apathetic* expression. We utilise these structural findings by reparameterising the expression subspace by the fourth-order moment tensor centred at the point of apathy. We propose a 3D face reconstruction method from single or multiple 2D projections by assuming an uncalibrated projective camera model. The non-linearity caused by the perspective projection can be neatly included into the model. The proposed algorithm separates person and expression subspaces convincingly, and enables flexible, natural modelling of expressions for a wide variety of human faces. Applying the method on independent faces showed that morphing between different persons and expressions can be performed without strong deformations.

Index Terms—Statistical shape model, tensor model, HOSVD, expression transfer, person transfer, 3D-reconstruction

1 Introduction

Human bodies and faces are able to perform a wide range of complex motions, and their variations in individual appearance and performance of movement make them difficult to capture and model. The topic of this work is the analysis of human faces represented by annotated, discrete 3D point sets of people who performed predefined emotions with varying strength. The variation of these point sets is then represented and characterised by a multiway array that naturally divides the data into shape, person, and expression modes that can be further decomposed by using conventional tensor decomposition techniques.

1.1 Related Work

Statistical shape spaces have been successfully used to model the complexities of complete human bodies and estimate accurate 3D-reconstructions from images [3], [4], [5], [6]. Constraints on known or unknown but constant [7], [8] limb lengths can increase the accuracy even further.

Using *principal component analysis (PCA)* to model face shape spaces is an old technique: *eigenfaces* are the principal components estimated from 2D-images showing static faces of many different persons [9]. A multiway array of 2D-images of different persons, expressions, viewpoints and illuminations was used to estimate a tensor factorisation [10] which can model these variations.

While these models are build upon image input, they are limited to 2D reconstruction and are not able to estimate a 3D shape. The first work proposing a factorisation approach to recover rigid 3D shapes from monocular 2D input sequences applies a low-rank constraint on the measurement array [11] and was extended to reconstruction of nonrigid 3D shapes from tracked 2D points using an orthographic projection in [12]. Since these algorithms assume a too restrictive [11] or a too general [12] model, they do not perform well to reconstruct 3D faces from image sequences if time-varying facial expression are observed throughout the sequences. Instead of relying on 2D image data solely, the authors in [13] use 3D-data obtained by a Laser Scanner, i.e. *dense* data. One PCA model is estimated to capture shape variations, another for texture variations. The *morphable model* can be used for 3D-reconstruction of human faces from 2D-images, for instance. Extensions with additionally varying expression mode have been proposed in [14], [15], [16]. The idea to have a known template was used in [17] to transfer expressions from one person to another in videos. It is based upon a morphable model and an explicitly known 3D target surface.

Instead of matrices, in [18] the authors propose to order the 3D face scans into higher order data representations, i.e. multiway array, and perform a tensor factorisation on it, to learn person, expression and viseme-related parameters directly from the 3D-point sets. The resulting model was used to perform face transfer on videos. Instead of directly using the 3D-point sets, a tensor model was estimated from the coefficients of localised wavelet transformations [19].

In [1], [2], we proposed a statistical shape model for a collection of human faces described as sets of 3D-points in different facial expressions. Similarly to the factorisation proposed in [18] and [19], the statistical shape model is based on representing the data in a multiway array and its factorisation by the higher-order singular value decomposition (HOSVD). The centre of the construction of models based on principle component analysis (PCA) or tensor factorisation [5], [20] is commonly the neutral expression. However there is controversy of the definition of the neutral

S. Graßhof is with Leibniz Universität Hannover and IT University of Copenhagen.

H. Ackermann, and J. Ostermann are with Leibniz Universität Hannover. E-mail: see http://www.tnt.uni-hannover.de/staff.

S. S. Brandt is with IT University of Copenhagen and University of Copenhagen. E-mail: sami.brandt@itu.dk.

The conference versions of this paper are [1] and [2].

This article was supported by German Research Foundation grant DFG RO 2497/12-2.

expression due to its ambiguous nature, while psychological studies have indicated it is not perceived as emotionless [21]. We found that the actual centre of the expression space, the *point of apathy* [1], is surprisingly not located in the origin. We thus tailor the expression analysis methods to be centred to the point of apathy. Using this new mode we are able to change the expression to *apathetic*, i.e. *neutral*, and hence perform a face neutralisation, or normalisation, which is a crucial step to improve face recognition [14].

Without regularisation, person or expression transfer fails, or is at least limited to small changes, thus prior works require strong constraints to allow for expression transfer between persons [22]. In [17], a separate 3D reconstruction of the target surface is computed in advance, and deviations from this template are penalised during expression transfer. This energy is nonlinear and non-convex, i.e. hard to optimise. In [18] and [19], no such penaliser is used, hence a transfer of expressions is limited to small changes.

A recent work based on deep neutral networks (NN) is [23], which relies on the Basel Model (BFM 2009) [24] extended by [25]. The authors present three different NNs for different tasks and, while they rely on landmarks during training, their approach is landmark-free during testing.

1.2 Contributions

In [1], [2] it was shown that these priors necessitate from a substructure the original data exhibits. After learning these structures, we propose using them to create an even more stable model, which implicitly penalises deviations from these substructures directly in parameter space. A contribution of this work is to use the formerly discovered apathetic facial expression to centre the face model and thereby lowering the number of parameters by encoding the emotion strength as norm of the emotion vector. Compared to previous works, we lowered the number of parameters needed to describe expressions from 25 [1], [2] to 6. Additionally, we propose neighbourhood sparsity constraints to favour solutions that are close to the training data, lowering the number of model parameters even further.

We present two different approaches to estimate the emotionless, apathetic facial expression, which was discovered as the root of all expressions in [1]: (1) We show that it can be directly computed from 3D face shapes of BU3DFE [26], and (2) by estimating it from a subspace of High-Order-SVD (HOSVD) based on 3D data of BU3DFE, as well as on 2D data of ADFES [27], which implies that the apathy mode is not an artefact of the applied factorisation of one dataset. Furthermore we automatically determine the penalty weights.

Given a tensor factorisation of the training data, unknown 3D-shapes can be inferred from images. This amounts to jointly estimating the mixing coefficients of the shape spaces as well as the parameters of the camera models which gave rise to the images. Just as in [2] we use uncalibrated projective cameras and show how the nonlinearity caused by the perspective projection can be linearised for the updated model.

The summary of our **contributions** is as follows:

- We propose a model based on a 4D tensor, which encodes the emotion and its strength in one parameter vector.

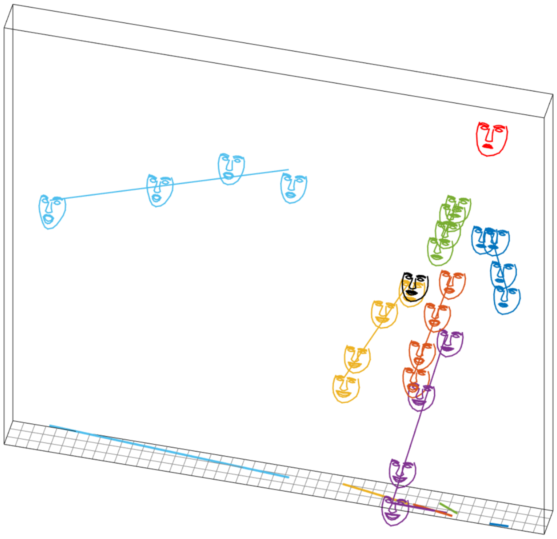


Figure 1. Expression space illustrated by the first three singular vectors of the expression dimension in the data tensor, i.e. the first three columns of $\mathbf{U}^{(3)}$. The levels of each of the six emotions form approximately linear trajectories meeting in a common vertex, *the point of apathy* (red), of which there was no explicit example in the training database. The space is oriented in the way that the stronger the emotion the further away from the apathy. (The same structure for the dense faces is illustrated in [1].) The colours represent the 7 emotions: neutral (*gray*), anger (*dark blue*), disgust (*orange*), fear (*yellow*), happiness (*violet*), sadness (*green*), surprise (*light blue*). (Please compare to Fig. 5(a).)

- The model is based on an apathy-centred expression space, and
- uses less parameters compared to the former model.
- We retrieve the apathetic face by directly estimating it from 3D shapes, as well as by HOSVD based on 3D data (BU3DFE [26]) and 2D data (ADFES [27]).
- Hyperparameters, i.e. weights of different error terms, are automatically determined.
- The performance of the proposed model is demonstrated on:
  - face synthesis,
  - face neutralisation,
  - person and expression transfer, and on
  - dense 3D reconstruction from sparse 2D landmarks.

The paper is organised as follows: The tensor factorisation model is introduced in Section 2. In Sec. 3 we show how to estimate the model parameters, assuming 3D points are provided. In Sec. 4 this approach is extended by incorporating a projective camera model, thereby enabling sparse 2D input. Experimental evaluations are presented in Section 5, where we present a validation of the apathy mode in Sec. 5.1–5.3 and emotion transfer in Sec. 5.4. In Sec. 5.5 the influence of the sparsity parameters is investigated, while in Sec. 5.6 the proposed model is used to estimate dense 3D reconstruction from sparse 2D landmarks. The conclusions are in Sec. 6.

1.3 Notation

In this paper we employ common notation, as follows: lower case italic letters define scalar values $s \in \mathbb{R}$, lower case bold letters define column vectors $\mathbf{v} \in \mathbb{R}^{N\times 1}$, and upper case bold letters are matrices $\mathbf{M} \in \mathbb{R}^{n\times m}$. For multilinear algebra we adopt the notation from [28], e.g. upper case slanted letters define tensors, hence a 3D tensor is $\mathcal{T} \in \mathbb{R}^{N\times P\times E}$ and the n-way product between a matrix $\mathbf{M}$ and the nth dimension of the tensor $\mathcal{T}$ is $\mathcal{T} \times_n \mathbf{M}$. This notation is used in [18], [19], [29] for shape modelling of faces.

2 Tensor Face Model

2.1 Three-Way Model

Assuming a set of 3D face scans of full correspondence is provided, the measurements are ordered in a data tensor $\mathcal{T}_{3,0} \in \mathbb{R}^{3N\times P\times E_{\text{tot}}}$, where N is the number of 3D vertices, P is the number of persons, and E_{tot} is the total number of expressions. All shapes are globally aligned by translation and rotation such that the top part of the nose is located at the origin for all shapes, while the individual scale is preserved. After that the mean face $\bar{\mathbf{f}}$ is calculated. Subtracting it from each shape gives the centred data tensor $\mathcal{T}_3 = \mathcal{T}_{3,0} - \overline{\mathcal{T}_3} \in \mathbb{R}^{3N\times P\times E_{\text{tot}}}$, $\overline{\mathcal{T}_3} = \bar{\mathbf{f}} \times_2 \mathbf{1}_P \times_3 \mathbf{1}_{E_{\text{tot}}}$ being the mean face tensor, where $\mathbf{1}_n \in \mathbb{R}^n$ defines a vector of length n, which only contains the value 1. The centred tensor can be decomposed by a Higher-Order-SVD (HOSVD) [28] as

$$\widehat{\mathcal{T}}_3 = \mathcal{S}_3 \times_1 \mathbf{U}_3^{(1)} \times_2 \mathbf{U}_3^{(2)} \times_3 \mathbf{U}_3^{(3)}, \qquad (1)$$

where $\mathcal{S}_3 \in \mathbb{R}^{3\widetilde{N}\times\widetilde{P}\times\widetilde{E}_{\text{tot}}}$ is the core tensor, and $\mathbf{U}_3^{(1)} \in \mathbb{R}^{3N\times 3\widetilde{N}}$, $\mathbf{U}_3^{(2)} \in \mathbb{R}^{P\times\widetilde{P}}$, $\mathbf{U}_3^{(3)} \in \mathbb{R}^{E_{\text{tot}}\times\widetilde{E}_{\text{tot}}}$ are orthogonal matrices[1], which consist of the singular vectors corresponding to the k-mode unfolded tensor, with $\widetilde{N} \leqslant N$, $\widetilde{P} \leqslant P$ and $\widetilde{E}_{\text{tot}} \leqslant E_{\text{tot}}$. Rewriting Eq. (1) for a face shape $\mathbf{f} \in \mathbb{R}^{3N}$, its approximation $\hat{\mathbf{f}}$ can be expressed as

$$\mathbf{f} \approx \hat{\mathbf{f}} = \bar{\mathbf{f}} + \mathcal{S}_3 \times_1 \mathbf{U}_3^{(1)} \times_2 \mathbf{w}_2^{\mathrm{T}} \times_3 \mathbf{w}_3^{\mathrm{T}}, \qquad (2)$$

where $\mathbf{w}_2 \in S^{\widetilde{P}}$ is the parameter vector for person and $\mathbf{w}_3 \in S^{\widetilde{E}_{\text{tot}}}$ of expression, and S^n denotes the n sphere[2]. This parameterisation was used in [18], [19], [29], and we refer to it as *base*.

In [1], [2], the following alternative parameterisation was demonstrated to be superior

$$\mathbf{f} \approx \hat{\mathbf{f}} = \bar{\mathbf{f}} + \mathcal{S}_3 \times_1 \mathbf{U}_3^{(1)} \times_2 \mathbf{p}_2^{\mathrm{T}}\mathbf{U}_3^{(2)} \times_3 \mathbf{p}_3^{\mathrm{T}}\mathbf{U}_3^{(3)}, \qquad (3)$$

where the updated model parameters $\mathbf{p}_2 \in \mathbb{R}^P$, $\mathbf{p}_3 \in \mathbb{R}^{E_{\text{tot}}}$ have the same or a higher dimension compared to the parameters $\mathbf{w}_2 \in \mathbb{R}^{\widetilde{P}}$, $\mathbf{w}_3 \in \mathbb{R}^{\widetilde{E}_{\text{tot}}}$ of Eq. (2), because $\widetilde{P} \leqslant P$, $\widetilde{E}_{\text{tot}} \leqslant E_{\text{tot}}$. However by using the latter parameterisation we will be able to constrain $\mathbf{w}_k$ on a lower dimensional manifold, whereas the original parameters are arbitrary. In [1], [2] we proposed additional constraints to employ the substructure of the subspaces and hence refer to it as *sub* in the remainder.

2.1.1 Substructure in Expression Space

A closer analysis on this three-way model reveals a special substructure in the expression space, as Fig. 1 illustrates. In other words, the expression space has a natural vertex at $\mathbf{w}_3 = \mathbf{w}_{\text{apathy}}$.

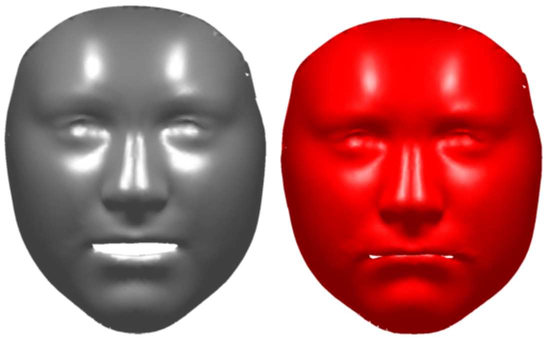


Figure 2. The neutral expression (left, grey) synthesised apathetic expression (right, red) are shown for one person of the database.

The HOSVD defined in Eq. (1) was calculated on a sparse $N = 83$ and a dense $N = 7308$ version of the data tensor. The first three dimensions of the third mode singular vectors of the expression spaces $\mathbf{U}^{(3)}$ are shown with corresponding 3D face shapes for the sparse model in Figure 1, which demonstrates the same structure as in the dense case [1]. For both models, it can be seen that all expressions lie on a planar substructure, in which four expression levels belonging to the same emotion can be approximated by one line. All these lines appear to intersect in one expression point at the top right, which is not part of the provided database and even more surprisingly, is not equal to the expression labelled as neutral. Inspecting the newly synthesised expression for several different persons, we labelled it as the *apathetic* facial expression. We defined it as such due to the impression that all facial muscles are completely relaxed, whereas this is not the case for the expression labelled as *neutral*. In Figure 2, we show the neutral expression in grey and the newly synthesised apathetic facial expression in red for one person. Please note that the latter does not exhibit an open mouth in Fig. 2, while the neutral face does. We consider this is a result of the fact that the database is build upon *posed* expressions, including some *neutral* expression with an open mouth. An important question is whether the apparent intersection of the emotion trajectories at the point of apathy as indicated by Fig. 1 is an effect of the higher-order tensor factorisation. In Sec. 5.1 two additional experiments are provided to support the claim of a unique point at which all emotion trajectories intersect. As soon as the vertex is found by regression (see Sec. 5.1.2), the apathetic faces can be collected to the apathy tensor

$$\mathcal{T}_{3,\text{apathy}} = \overline{\mathcal{T}_3} + \mathcal{S}_3 \times_1 \mathbf{U}_3^{(1)} \times_2 \mathbf{U}_3^{(2)} \times_3 \left(\mathbf{1}_{E_{\text{tot}}} \mathbf{w}_{\text{apathy}}^{\mathrm{T}}\right), \qquad (4)$$

where $\mathbf{1}_{E_{\text{tot}}} = (1, \ldots, 1)^{\mathrm{T}} \in \mathbb{R}^{E_{\text{tot}}}$. This tensor collects the apathetic facial expressions varying between persons. The mean face of all apathetic faces is defined as $\bar{\mathbf{f}}_{\text{apathy}}$, whereas its 3-dimensional tensor representation is denoted as $\overline{\mathcal{T}}_{3,\text{apathy}}$.

2.2 Four-Way Model

To exploit the revealed structure found in the expression space, the natural way is to centre the tensor into the point of apathy. Moreover, it is natural to fold the emotion strength to its own dimension in the data tensor to separate it from the emotion, i.e. separating the expression into emotion

[1] There is, however, a sign ambiguity of the singular vectors since any left–right singular vector pair $(\mathbf{u}, \mathbf{v})$ of a matrix can be equivalently replaced by $(-\mathbf{u}, -\mathbf{v})$. To resolve this ambiguity, we select the sign for the singular vectors so that the first element of each left singular vector is always non-negative.

[2] $S^n = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| \leqslant 1\}$.

and its strength. In this way, the expression trajectories ideally form one-dimensional linear subspaces where the zero strength would correspond to the point of apathy while all expect the most dominant strength-mode singular vectors can be truncated.

Formally, we therefore represent the original data, and the apathy tensor, by folding them into $3N \times P \times S \times E$, four-way data tensor $\mathcal{T}_{4,0}$ and mean apathy tensor $\overline{\mathcal{T}}_{4,\text{apathy}}$, where S refers to the emotion strength and E to the emotion. Please note that the neutral expression is skipped for this reordering for two reasons: First, there is only one possible emotion strength for neutral provided, and second it is a non consistent expression, but is performed very inconsistently with varying appearances, e.g. some with open other with closed mouth. Let $\mathcal{T}_4 = \mathcal{T}_{4,0} - \overline{\mathcal{T}}_{4,\text{apathy}}$ be the apathy centred tensor that is approximated as

$$\widehat{\mathcal{T}}_4 = \mathcal{S}_4 \times_1 \mathbf{U}_4^{(1)} \times_2 \mathbf{U}_4^{(2)} \times_3 \mathbf{U}_4^{(3)} \times_4 \mathbf{U}_4^{(4)}, \tag{5}$$

where $\mathcal{S}_4 \in \mathbb{R}^{3\widetilde{N} \times \tilde{P} \times \tilde{S} \times \tilde{E}}$ is the core tensor, and $\mathbf{U}_4^{(1)} \in \mathbb{R}^{3N \times 3\widetilde{N}}$, $\mathbf{U}_4^{(2)} \in \mathbb{R}^{P \times \tilde{P}}$, $\mathbf{U}_4^{(3)} \in \mathbb{R}^{S \times \tilde{S}}$, $\mathbf{U}_4^{(4)} \in \mathbb{R}^{E \times \tilde{E}}$ with $\widetilde{N} \leqslant N$, $\widetilde{P} \leqslant P$, $\widetilde{S} \leqslant S$, and $\widetilde{E} \leqslant E$. Similarly as above, the faces can be approximated by the four way model as

$$\mathbf{f} \approx \overline{\mathbf{f}}_{\text{apathy}} + \mathcal{S}_4 \times_1 \mathbf{U}_4^{(1)} \times_2 \mathbf{w}_2^{\mathrm{T}} \times_3 \mathbf{w}_3^{\mathrm{T}} \times_4 \mathbf{w}_4^{\mathrm{T}}, \tag{6}$$

where $\mathbf{w}_2 \in S^{\tilde{P}}$ is the parameter vector for person, $\mathbf{w}_3 \in S^{\tilde{S}}$ of strength, and $\mathbf{w}_4 \in S^{\tilde{E}}$ for emotion.

Assuming then that the expressions are one-dimensional linear subspaces centred at the apathetic faces implies that $\widetilde{S} = 1$, hence $\mathbf{U}_4^{(3)} \in \mathbb{R}^{S \times 1}$ and $\mathbf{w}_3 \equiv w_3$ is a scalar, the *emotion strength* parameter. In this case the core tensor can be truncated and we may define $\widetilde{\mathcal{S}}_4$ as the corresponding $3N \times P \times E$, which is obtained by trivially unfolding the singleton strength dimension that yields

$$\mathbf{f} \approx \widehat{\mathbf{f}} = \overline{\mathbf{f}}_{\text{apathy}} + \widetilde{\mathcal{S}}_4 \times_1 \mathbf{U}_4^{(1)} \times_2 \mathbf{w}_2^{\mathrm{T}} \times_3 \mathbf{w}_{34}^{\mathrm{T}}, \tag{7}$$

where $\mathbf{w}_{34} \equiv w_3 \mathbf{w}_4$, hence, the expression parameter vector is modulated by the scalar strength parameter.

Transferring this to the latest model parameterisation of Eq. (3) in consequence leads to

$$\begin{aligned} \widehat{\mathbf{f}} &= \overline{\mathbf{f}}_{\text{apathy}} + \dots \\ &\quad \mathcal{S}_4 \times_1 \mathbf{U}_4^{(1)} \times_2 \mathbf{p}_2^{\mathrm{T}} \mathbf{U}_4^{(2)} \times_3 \mathbf{p}_3^{\mathrm{T}} \mathbf{U}_4^{(3)} \times_4 \mathbf{p}_4^{\mathrm{T}} \mathbf{U}_4^{(4)}, && (8) \\ &= \overline{\mathbf{f}}_{\text{apathy}} + \widetilde{\mathcal{S}}_4 \times_1 \mathbf{U}_4^{(1)} \times_2 \mathbf{p}_2^{\mathrm{T}} \mathbf{U}_4^{(2)} \times_3 \underbrace{w_3 \mathbf{p}_4^{\mathrm{T}}}_{\mathbf{p}_{34}^{\mathrm{T}}} \mathbf{U}_4^{(4)}, && (9) \end{aligned}$$

where $\|\mathbf{p}_{34}\| = w_3$ and $\mathbf{p}_4 = \mathbf{p}_{34}/w_3$.

2.3 Tensor Model vs. PCA

The HOSVD model of Eq. (5) is connected to PCA, see also [30], as follows. Let $\mathbf{T}_4^{(n)}$ denote the n-mode matrix unfolding of $\mathcal{T}_4$, and $\mathbf{T}_{4,\text{apathy}}^{(k)}$ the n-mode unfolding of $\overline{\mathcal{T}}_{4,\text{apathy}}$. Let us define

$$\begin{aligned} \mathbf{R}^{(1)} &= \frac{1}{PSE}\left(\mathbf{T}_{4,0}^{(1)} - \mathbf{T}_{4,\text{apathy}}^{(1)}\right)\left(\mathbf{T}_{4,0}^{(1)} - \mathbf{T}_{4,\text{apathy}}^{(1)}\right)^{\mathrm{T}} \\ &= \frac{1}{PSE}\mathbf{T}_4^{(1)}{\mathbf{T}_4^{(1)}}^{\mathrm{T}} && (10) \\ &= \frac{1}{PSE}\mathbf{U}_4^{(1)}\mathbf{S}_4^{(1)}{\mathbf{V}_4^{(1)}}^{\mathrm{T}}\mathbf{V}_4^{(1)}{\mathbf{S}_4^{(1)}}^{\mathrm{T}}{\mathbf{U}_4^{(1)}}^{\mathrm{T}} && (11) \\ &= \mathbf{U}_4^{(1)}\mathbf{\Lambda}^{(1)}{\mathbf{U}_4^{(1)}}^{\mathrm{T}}, && (12) \end{aligned}$$

where $\mathbf{\Lambda}^{(1)} = \frac{1}{PSE}\mathbf{S}_4^{(1)}{\mathbf{S}_4^{(1)}}^{\mathrm{T}}$. This means that the matrix $\mathbf{R}^{(1)}$ corresponds to the sample covariance matrix $\mathbf{C}^{(1)}$ of all faces with the difference that the faces are centered at the apathetic face where the covariance is centred at the mean face. In other words, where the 1-mode principal components are the eigenvectors of the covariance matrix $\mathbf{C}^{(1)}$, the 1-mode singular vectors are the eigenvectors of $\mathbf{R}^{(1)}$. Similarly, for the person or 2-mode matrix

$$\begin{aligned} \mathbf{R}^{(2)} &= \frac{1}{SEN}\left(\mathbf{T}_{4,0}^{(2)} - \mathbf{T}_{4,\text{apathy}}^{(2)}\right)\left(\mathbf{T}_{4,0}^{(2)} - \mathbf{T}_{4,\text{apathy}}^{(2)}\right)^{\mathrm{T}} \\ &= \mathbf{U}_4^{(2)}\mathbf{\Lambda}^{(2)}{\mathbf{U}_4^{(2)}}^{\mathrm{T}}, && (13) \end{aligned}$$

where $\mathbf{\Lambda}^{(2)} = \frac{1}{SEN}\mathbf{S}_4^{(2)}{\mathbf{S}_4^{(2)}}^{\mathrm{T}}$. the 2-mode matrix unfolding of the tensor contains the matched measurements, centred at the apathetic face, for all the people in each column. Therefore the 2-mode singular vectors describe the directions in the people space showing the largest spread centred at the apathetic face measurements. For the strength or 3-mode,

$$\begin{aligned} \mathbf{R}^{(3)} &= \frac{1}{ENP}\left(\mathbf{T}_{4,0}^{(3)} - \mathbf{T}_{4,\text{apathy}}^{(3)}\right)\left(\mathbf{T}_{4,0}^{(3)} - \mathbf{T}_{4,\text{apathy}}^{(3)}\right)^{\mathrm{T}} \\ &= \mathbf{U}_4^{(3)}\mathbf{\Lambda}^{(3)}{\mathbf{U}_4^{(3)}}^{\mathrm{T}}, && (14) \end{aligned}$$

where $\mathbf{\Lambda}^{(3)} = \frac{1}{ENP}\mathbf{S}_4^{(3)}{\mathbf{S}_4^{(3)}}^{\mathrm{T}}$. The 3-mode matrix unfolding contains the matched measurement in the function of the emotion strength, with respect to the apathetic expression, in each column. Since the point of apathy is the origin and the strength increases when moving away from it, it is easy to see how the rank-1 approximation works in the strength space: in the approximation the columns will linearly dependent, i.e., equivalent up to a scalar multiplier. Finally, the emotion or the 4-mode

$$\begin{aligned} \mathbf{R}^{(4)} &= \frac{1}{NPS}\left(\mathbf{T}_{4,0}^{(4)} - \mathbf{T}_{4,\text{apathy}}^{(4)}\right)\left(\mathbf{T}_{4,0}^{(4)} - \mathbf{T}_{4,\text{apathy}}^{(4)}\right)^{\mathrm{T}} \\ &= \mathbf{U}_4^{(4)}\mathbf{\Lambda}^{(4)}{\mathbf{U}_4^{(4)}}^{\mathrm{T}}, && (15) \end{aligned}$$

where $\mathbf{\Lambda}^{(4)} = \frac{1}{NPS}\mathbf{S}_4^{(4)}{\mathbf{S}_4^{(4)}}^{\mathrm{T}}$. The 4-mode matrix unfolding contains all the expressions, in the form of matched measurements, in each column. Therefore, the 4-mode singular vectors describe the directions of largest spread in the expression space relative to the apathetic expression.

3 Estimation of Tensor Model Parameters

In real applications, such as approximation or expression transfer, see Fig. 3, one needs to fit the underlying model to a novel face, while the model may be likewise updated by additional training data. The relationship of the models

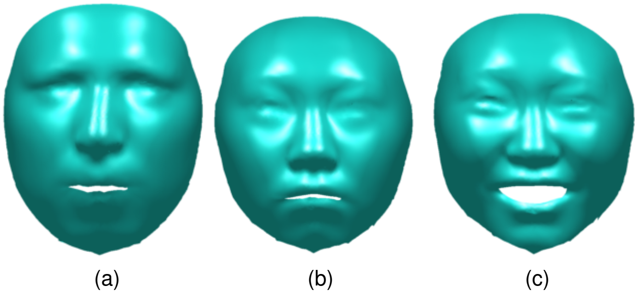


Figure 3. Transfer *anger* from one person to another: (a) person A in emotion anger, (b) person B with estimated emotion anger, (c) person B, in emotion happy.

$$\begin{tikzcd}
\cdots \arrow[r, "D_n"'] & S_n \arrow[r, "D_{n+1}"'] \arrow[d, "\theta"'] & S_{n+1} \arrow[r, "D_{n+2}"'] \arrow[d, "\theta"'] & \cdots \\
 & \mathbf{f}|\theta, S_n & \mathbf{f}|\theta, S_{n+1} &
\end{tikzcd}$$

Figure 4. Diagram characterising the statistical model used in this paper. The state S_n refers to the HOSVD model $\mathcal{S}_4, \mathbf{U}_4^{(1)}, \mathbf{U}_4^{(2)}, \mathbf{U}_4^{(3)}, \mathbf{U}_4^{(4)}$ trained by the data D_n. The model also includes the parameters $\widetilde{N}, \widetilde{P}, \widetilde{S}, \widetilde{E}$. The novel face $\mathbf{f}$, modelled by the parameters θ, depends on the current state. The state, or the HOSVD model, may be updated by adding more training data D_{n+1}.

is characterised by the diagram in Fig. 4. In the following section we look at the case where, the expression and person parameters are unknown and need to be estimated for a novel face. In Fig. 3 we show that we can transfer an expression from one person to another using the estimated parameters from our latest model.

3.1 Statistical Objective

Given a novel 3D face $\mathbf{f}$, our objective is to compute the maximum posterior (MAP) estimate for the model parameters $\theta = \{\mathbf{w}_2, \mathbf{w}_3, \mathbf{w}_4\}$ of the model Eq. (6), by maximising the posterior

$$p(\theta|\mathbf{f}, S_n) = p(\mathbf{f}|\theta, S_n)p(\theta|\alpha, S_n) \tag{16}$$

where $p(\mathbf{f}|\theta, S_n)$ is the likelihood and $p(\theta|\alpha, S_n)$ is the prior, S_n is the state (c.f. Fig. 4), and $\alpha = \{\alpha_{\mathrm{P}}, \alpha_{\mathrm{E}}\}$ represents hyperparameters. For the likelihood part, we assume i.i.d. Gaussian noise. From now on, let us assume the truncated model Eq. (7), where $\widetilde{S} = 1$. Across the person and expression dimensions, the prior is independent, or,

$$p(\mathbf{w}_2, \mathbf{w}_3, \mathbf{w}_4|\alpha, S_n) = p(\mathbf{w}_2|\alpha, S_n)p(\mathbf{w}_{34}|\alpha, S_n). \tag{17}$$

For the person mode, we assume a smooth, piece-wise linear person parameter manifold as follows. We assume the novel person parameters are a convex combination of only α_{P} close training people in the database while a Gaussian prior is assumed for the person vectors. Formally,

$$p(\mathbf{w}_2|\alpha, S_n) = p(\mathbf{w}_2|\mathcal{W}_{\mathrm{P}}, \alpha, S_n) \propto \begin{cases} g_{\sigma_{\mathrm{P}}}(\mathbf{w}_2), & \mathbf{w}_2 \in \mathcal{W}_{\mathrm{P}} \\ 0, & \mathbf{w}_2 \notin \mathcal{W}_{\mathrm{P}} \end{cases} \tag{18}$$

where $g_{\sigma_{\mathrm{P}}}$ is the zero mean, isotropic Gaussian probability density function (pdf) with covariance matrix $\sigma_{\mathrm{P}}\mathbf{I}_{\tilde{P}}$, and

$$\begin{aligned} \mathcal{W}_{\mathrm{P}} = \Bigg\{ \mathbf{w}_2 \in \mathbb{R}^{\tilde{P}} \Bigg| \, & \mathbf{w}_2 = {\mathbf{U}_4^{(2)}}^{\mathrm{T}} \mathbf{p}_2, \text{ where } \mathbf{p}_2 \in \mathbb{R}^P \wedge \\ & \mathbf{p}_2^{\mathrm{T}} \mathbf{1}_P = 1 \wedge \mathbf{p}_2 \geqslant \mathbf{0} \wedge \|\mathbf{p}_2\|_0 = \alpha_{\mathrm{P}}, \text{ and the non-} \\ & \text{zero elements indicate the } \alpha_{\mathrm{P}}\text{-neighbourhood} \\ & \text{among the rows in } \mathbf{U}_4^{(2)} \Bigg\}. \end{aligned} \tag{19}$$

Here, $\mathbf{p}_2$ is the hyper parameter vector of person related parameters.

For the expression mode, we assume that novel expression parameters $\mathbf{w}_{34}$ are a convex combination of only α_{E} close training expressions in the database. In addition, we have a Gaussian prior for the expression vector centred at the point of apathy. Thus we obtain,

$$p(\mathbf{w}_{34}|\alpha, S_n) = p(\mathbf{w}_{34}|\mathcal{W}_{\mathrm{E}}, \alpha, S_n) \propto \begin{cases} g_{\sigma_{\mathrm{E}}}(\mathbf{w}_{34}), & \mathbf{w}_{34} \in \mathcal{W}_{\mathrm{E}} \\ 0, & \mathbf{w}_{34} \notin \mathcal{W}_{\mathrm{E}} \end{cases} \tag{20}$$

where $g_{\sigma_{\mathrm{E}}}$ is the zero mean, isotropic Gaussian probability density function (pdf) with covariance matrix $\sigma_{\mathrm{E}}\mathbf{I}_{\tilde{E}}$, and

$$\begin{aligned} \mathcal{W}_{\mathrm{E}} = \Bigg\{ \mathbf{w}_{34} \in \mathbb{R}^{\tilde{E}} \Bigg| \, & \mathbf{w}_{34} = {\mathbf{U}_4^{(4)}}^{\mathrm{T}} \mathbf{p}_{34}, \text{ where } \mathbf{p}_{34}, \\ & \mathbf{p}_4 \in \mathbb{R}^E \wedge \mathbf{p}_4^{\mathrm{T}} \mathbf{1}_P = 1 \wedge \mathbf{p}_4 \geqslant \mathbf{0} \wedge (\|\mathbf{p}_{34}\|_0 = \alpha_{\mathrm{E}} \vee \\ & \|\mathbf{p}_{34}\|_0 = 0), \|\mathbf{p}_{34}\|\mathbf{p}_4 = \mathbf{p}_{34}, \text{ and the non-} \\ & \text{zero elements indicate the } \alpha_{\mathrm{E}}\text{-neighbourhood} \\ & \text{among the rows in } \mathbf{U}_4^{(4)} \Bigg\}. \end{aligned} \tag{21}$$

Here, $\mathbf{p}_{34}$ is the hyper parameter vector of expression related parameters. The statistical model implies the following MAP estimation problem

$$\min_{\mathbf{p}_2, \mathbf{p}_{34}} \frac{1}{2}\|\widehat{\mathbf{f}} - \mathbf{f}\|_2^2 + \frac{\lambda_2}{2}\|{\mathbf{U}_4^{(2)}}^{\mathrm{T}}\mathbf{p}_2\|_2^2 + \frac{\lambda_{34}}{2}\|{\mathbf{U}_4^{(4)}}^{\mathrm{T}}\mathbf{p}_{34}\|_2^2, \tag{22}$$

subject to

$$\begin{aligned} \mathbf{p}_2 \geqslant \mathbf{0}, \quad \mathbf{p}_4 \geqslant \mathbf{0}, \quad \mathbf{p}_2^{\mathrm{T}}\mathbf{1}_P = 1, \quad \mathbf{p}_4^{\mathrm{T}}\mathbf{1}_E = 1, \\ \|\mathbf{p}_2\|_0 = \alpha_{\mathrm{P}}, \quad \|\mathbf{p}_{34}\|_0 = \alpha_{\mathrm{E}}, \end{aligned} \tag{23}$$

where the non-zero elements indicate the α_{P}-neighbourhood among row vectors in $\mathbf{U}_4^{(2)}$, and α_{E}-neighbourhood among the row vectors in $\mathbf{U}_4^{(4)}$, respectively, and $\widehat{\mathbf{f}}$ as defined by Eq. (9). The numerical optimisation of Eq. (22) is described in the following section.

3.2 Numerical Optimisation

In this section, we will first describe how to rewrite the tensor-product into a matrix-vector product, and then describe the optimisation for all considered models.

3.2.1 Linear Matrix-vector Model Representation

All presented models require tensor products to compute a face shape, which prevents a closed-form solution for both parameters. However, by rewriting the previous tensor model such that it is linear in one model parameter in the matrix-vector notation, it allows us to use alternating least-squares (ALS) to estimate the parameter vectors.

Let us assume that expression parameter vector $\mathbf{w}_3^{\mathrm{T}}$ is fixed, and reorder the elements of Eq. (7) to

$$\widehat{\mathbf{f}} - \overline{\mathbf{f}}_{\mathrm{apathy}} = \widetilde{\mathcal{S}}_4 \times_1 \mathbf{U}_4^{(1)} \times_3 \mathbf{w}_{34}^{\mathrm{T}} \times_2 \mathbf{w}_2^{\mathrm{T}}. \tag{24}$$

The tensor $\widetilde{\mathcal{S}}_4 \times_1 \mathbf{U}^{(1)} \times_3 \mathbf{w}_{34}^{\mathrm{T}} \in \mathbb{R}^{3N \times \widetilde{P} \times 1}$ can be trivially flattened into a $3N \times \widetilde{P}$ matrix $\mathbf{M}_2$ as

$$\mathbf{M}_2\, \mathbf{w}_2 \equiv \widetilde{\mathcal{S}}_4 \times_1 \mathbf{U}_4^{(1)} \times_3 \mathbf{w}_{34}^{\mathrm{T}} \times_2 \mathbf{w}_2^{\mathrm{T}} \tag{25}$$

thus

$$\widehat{\mathbf{f}} - \overline{\mathbf{f}}_{\mathrm{apathy}} = \mathbf{M}_2 {\mathbf{U}_4^{(2)}}^{\mathrm{T}} \mathbf{p}_2, \tag{26}$$

hence, the difference is linear in $\mathbf{p}_2$. Accordingly for $\mathbf{p}_{34}$:

$$\widehat{\mathbf{f}} - \overline{\mathbf{f}}_{\mathrm{apathy}} = \widetilde{\mathcal{S}}_4 \times_1 \mathbf{U}_4^{(1)} \times_2 \mathbf{w}_2^{\mathrm{T}} \times_3 \mathbf{w}_{34}^{\mathrm{T}}, \tag{27}$$

where the elements of $\widetilde{\mathcal{S}}_4 \times_1 \mathbf{U}_4^{(1)} \times_2 \mathbf{w}_2^{\mathrm{T}} \in \mathbb{R}^{3N \times 1 \times \tilde{E}}$ can be sorted into the matrix $\mathbf{M}_{34} \in \mathbb{R}^{3N \times \tilde{E}}$, leading to

$$\widehat{\mathbf{f}} - \overline{\mathbf{f}}_{\mathrm{apathy}} = \mathbf{M}_{34} {\mathbf{U}_4^{(4)}}^{\mathrm{T}} \mathbf{p}_{34}. \tag{28}$$

3.2.2 Optimisation

Taking only the terms depending on $\mathbf{p}_2$ from Eq. (22) yields the energy functional

$$\begin{aligned} E_2(\mathbf{p}_2) &= \frac{1}{2}\|\widehat{\mathbf{f}} - \mathbf{f}\|_2^2 + \frac{\lambda_2}{2}\|{\mathbf{U}_4^{(2)}}^{\mathrm{T}} \mathbf{p}_2\|_2^2 + C & (29)\\ &= \frac{1}{2}\mathbf{p}_2^{\mathrm{T}} \mathbf{Q}_2 \mathbf{p}_2 + \mathbf{b}_2^{\mathrm{T}} \mathbf{p}_2 + C, & (30) \end{aligned}$$

where C refers to the summand in Eq. (22) not containing $\mathbf{p}_2$, $\mathbf{Q}_2 = \mathbf{U}_4^{(2)} \left(\mathbf{M}_2^{\mathrm{T}} \mathbf{M}_2 + \lambda_2 \mathbf{I}\right) {\mathbf{U}_4^{(2)}}^{\mathrm{T}}$ and $\mathbf{b}_2 = -\mathbf{U}_4^{(2)} \mathbf{M}_2^{\mathrm{T}} (\mathbf{f} - \overline{\mathbf{f}}_{\mathrm{apathy}})$. We thus have the minimisation problem

$$\min_{\mathbf{p}_2} \frac{1}{2}\mathbf{p}_2^{\mathrm{T}} \mathbf{Q}_2 \mathbf{p}_2 + \mathbf{b}_2^{\mathrm{T}} \mathbf{p}_2 \tag{31}$$

subject to $\mathbf{p}_2 \geqslant \mathbf{0}$, $\mathbf{p}_2^{\mathrm{T}} \mathbf{1}_P = 1$, $\|\mathbf{p}_2\|_0 = \alpha_{\mathrm{P}}$, where the non-zero elements form the α_{P}-neighbourhood among the row vectors in $\mathbf{U}_4^{(2)}$.

To form the neighbourhood sparsity constraints, we find the α_{P}-nearest neighbours for each row vector in $\mathbf{U}_4^{(2)}$. The minimisation is performed separately over all these neighbourhoods, i.e., we define the projection $\mathbf{P}_2^i$ as the sparse matrix whose element $p_{jk} = 1$ iff the row j in $\mathbf{U}_2^{(2)}$ is the $k\mathrm{th}$ nearest neighbour of the point i and $k = 1, 2, \ldots, \alpha_{\mathrm{P}}$, and $p_{jk} = 0$, otherwise. Noting that $\mathbf{q}_2$ equals ${\mathbf{P}_2^i}^{\mathrm{T}} \mathbf{p}_2$, we may write the minimisation Eq. (31) in the equivalent form

$$\min_i \min_{\mathbf{q}_2} \frac{1}{2}\mathbf{q}_2^{\mathrm{T}} \mathbf{Q}_2^i \mathbf{q}_2 + {\mathbf{b}_2^i}^{\mathrm{T}} \mathbf{q}_2, \tag{32}$$

subject to $\mathbf{q}_2 \geqslant \mathbf{0}, \quad \mathbf{q}_2^{\mathrm{T}} \mathbf{1}_{\alpha_{\mathrm{P}}} = 1$, where
$\mathbf{Q}_2^i = {\mathbf{P}_2^i}^{\mathrm{T}} \mathbf{U}_4^{(2)} \left(\mathbf{M}_2^{\mathrm{T}} \mathbf{M}_2 + \lambda_2 \mathbf{I}\right) {\mathbf{U}_4^{(2)}}^{\mathrm{T}} \mathbf{P}_2^i$ and
$\mathbf{b}_2^i = -{\mathbf{P}_2^i}^{\mathrm{T}} \mathbf{U}_4^{(2)} \mathbf{M}_2^{\mathrm{T}} (\mathbf{f} - \overline{\mathbf{f}}_{\mathrm{apathy}})$. Similarly, considering the minimisation of Eq. (22) over $\mathbf{p}_{34}$ yields the minimisation problem

Algorithm 1 ALS for sparse parameter recovery

```
Input: 3D points of a face shape $\mathbf{f}$
• Initialisation
    – Global rigid alignment of 3D input and face model
    – Initialise expression parameter vector to the mean expression $\widehat{\mathbf{p}}_{34} := \frac{1}{E}(1, \ldots, 1)^{\mathrm{T}}$
• Model Parameter Estimation
    Repeat until convergence:
    – Given current expression parameter vector $\widehat{\mathbf{p}}_{34}$, estimate $\mathbf{p}_2$ by minimising Eq. (32), and set $\widehat{\mathbf{p}}_2 = \mathbf{P}_2^{\widehat{i}} \widehat{\mathbf{q}}_2$.
    – Given the current person parameter $\widehat{\mathbf{p}}_2$, estimate $\mathbf{p}_{34}$ by minimising Eq. (33), and set $\widehat{\mathbf{p}}_{34} = \mathbf{P}_{34}^{\widehat{i'}} \widehat{\mathbf{q}}_{34}$.
    – While solving Eq. (32) and Eq. (33), start with one $\lambda_k$ and adapt as described in Sec.3.2.3.
Output: $\widehat{\mathbf{p}}_2$, $\widehat{\mathbf{p}}_{34}$, $\widehat{\mathbf{f}} \in \mathbb{R}^{3N}$ from Eq. (9)
```

$$\min_{i'} \min_{\mathbf{q}_{34}} \frac{1}{2}\mathbf{q}_{34}^{\mathrm{T}} \mathbf{Q}_{34}^{i'} \mathbf{q}_{34} + {\mathbf{b}_{34}^{i'}}^{\mathrm{T}} \mathbf{q}_{34}, \tag{33}$$

subject to $\mathbf{q}_{34} \geqslant \mathbf{0}$, where $\mathbf{b}_{34}^{i'} = -{\mathbf{P}_4^{i'}}^{\mathrm{T}} \mathbf{U}_4^{(4)} \mathbf{M}_{34}^{\mathrm{T}} (\mathbf{f} - \overline{\mathbf{f}}_{\mathrm{apathy}})$ and $\mathbf{Q}_{34}^{i'} = {\mathbf{P}_4^{i'}}^{\mathrm{T}} \mathbf{U}_4^{(4)} \left(\mathbf{M}_{34}^{\mathrm{T}} \mathbf{M}_{34} + \lambda_{34} \mathbf{I}\right) {\mathbf{U}_4^{(4)}}^{\mathrm{T}} \mathbf{P}_4^{i'}$.

The minimisation problems Eq. (32) and Eq. (33) with the constraints can be performed using an interior-point convex quadratic programming such as `quadprog` in Matlab. The method is now complete and summarised in Algorithm 1. The iteration is stopped if the maximum number of iterations is reached, or if any of the following values is below a predefined threshold: the MSE (mean squared error) between input and approximated shape, or the change of proceeding errors, likewise.

3.2.3 Automatic Penalty Weights

The optimisation functions in Eq. (32) and Eq. (33) contain a penalty weight parameter λ, encoded in the matrix $\mathbf{Q}$, which is commonly selected manually. If the individual constraints and indices are ignored, the previously described minimisation problems share the following structure

$$\min_{\mathbf{q}} \frac{1}{2}\mathbf{q}^{\mathrm{T}} \mathbf{Q}_\lambda \mathbf{q} + \mathbf{b}^{\mathrm{T}} \mathbf{q}, \tag{34}$$

where $\mathbf{Q}_\lambda$ refers to a matrix, which depends on λ:

$$\mathbf{Q}_\lambda = \mathbf{P}^{\mathrm{T}} \mathbf{U} \left(\mathbf{M}^{\mathrm{T}} \mathbf{M} + \lambda \mathbf{I}\right) \mathbf{U}^{\mathrm{T}} \mathbf{P}. \tag{35}$$

To determine the best parameter λ a linesearch procedure is applied, i.e. given the current estimate of the parameter vector $\mathbf{q}$ and the corresponding constraints, the selected λ gives a local minimum of the former optimisation function.

4 Recovering dense 3D Faces from sparse 2D Landmarks

Previously we assumed that 3D points of a face are given to approximate a 3D face shape. However, most commonly only 2D projections of the faces are available for which sparse 2D landmarks can be computed. From these a dense 3D face can be reconstructed by the proposed multilinear model. Comparing 3D model points to 2D image coordinates requires a mapping from 3D to 2D by a reasonable camera model, which we define and use in the following.

4.1 Projective Camera Model

An often used camera model in computer vision tasks, e.g. 3D reconstruction, is the weak-perspective camera model [31], [32]. To account for perspective distortion, we choose a projective camera model and assume that pixels are square elements on the image sensor, which is given by most consumer cameras. A 3D point $\mathbf{x} \in \mathbb{R}^3$ is mapped to a 2D location $\mathbf{u} \in \mathbb{R}^2$ as

$$\begin{aligned}\tilde{\mathbf{u}} &= (\tilde{u}_x, \tilde{u}_y, \tilde{u}_z)^{\mathrm{T}} = \mathbf{K}\left(\mathbf{R}\mathbf{x}+\mathbf{t}\right),\\ \mathbf{u} &= (u_x, u_y)^{\mathrm{T}} = (\tilde{u}_x/\tilde{u}_z, \tilde{u}_y/\tilde{u}_z)^{\mathrm{T}},\end{aligned} \tag{36}$$

where $\mathbf{K} := \begin{pmatrix} fs_x & 0 & p_x \\ 0 & fs_y & p_y \\ 0 & 0 & 1 \end{pmatrix} \in \mathbb{R}^{3\times 3}$, (p_x, p_y) is the principal point, $f \in \mathbb{R}^+$ is the focal length, $\mathbf{R}$ is a 3D rotation matrix and $\mathbf{t}$ is a 3D translation vector.

4.2 Camera Projected 3D Face Tensor Model

Assuming that the camera parameters are provided, the model parameters of the proposed model of Eq. (9) can be estimated linearly [2].

Given n 2D landmarks corresponding to a subset of the N 3D model vertices, the 3D model point $\widehat{\mathbf{f}}_i$ is obtained from $\widehat{\mathbf{f}}$ by selecting specific rows, and accordingly $\mathbf{M}_{2,i}$ is obtained from $\mathbf{M}_2$. In the following the mean apathetic face is referred to as $\bar{\mathbf{f}}$. Using this notation, a 3D face point, generated by the model, can be defined linearly in the parameters $\mathbf{p}_2$ and $\mathbf{p}_{34}$ so that $\widehat{\mathbf{f}}_i = \mathbf{M}_{2,i}{\mathbf{U}_4^{(2)}}^{\mathrm{T}}\mathbf{p}_2 + \bar{\mathbf{f}}_i$ and $\widehat{\mathbf{f}}_i = \mathbf{M}_{3,i}{\mathbf{U}_4^{(4)}}^{\mathrm{T}}\mathbf{p}_{34} + \bar{\mathbf{f}}_i$. To distinguish between 3D and 2D, the faces will be referred to as $\mathbf{f}^{\mathrm{3D}}$ and $\mathbf{f}^{\mathrm{2D}}$ in the following.

4.2.1 Projective Face Tensor Model

Assuming camera parameters for the projective camera are provided, a 3D point $\mathbf{f}_i^{\mathrm{3D}}$ is mapped to its corresponding 2D point $\mathbf{f}_i^{\mathrm{2D}}$ by Eq. (36), or,

$$\tilde{\mathbf{u}}_i = (\tilde{u}_{i,x}, \tilde{u}_{i,y}, \tilde{u}_{i,z})^{\mathrm{T}} = \mathbf{K}\left(\mathbf{R}\mathbf{f}_i^{\mathrm{3D}} + \mathbf{t}\right), \tag{37}$$

$$\mathbf{f}_i^{\mathrm{2D}} = (\tilde{u}_{i,x}/\tilde{u}_{i,z}, \tilde{u}_{i,y}/\tilde{u}_{i,z})^{\mathrm{T}}. \tag{38}$$

Thus, 2D points are not linearly related to their 3D counterparts if a projective camera model is employed. We therefore rewrite Eq. (38) component-wise to retrieve a form which is linear in $\mathbf{p}_2$ in [2]. Similarly to [33], [34], the x component the 2D face shape Eq. (38) can be rewritten

$$\begin{aligned} f_{i,x}^{\mathrm{2D}} &= \tilde{u}_{i,x}/\tilde{u}_{i,z} \\ \Leftrightarrow f_{i,x}^{\mathrm{2D}}\,\tilde{u}_{i,z} &= \tilde{u}_{i,x} \\ \Leftrightarrow f_{i,x}^{\mathrm{2D}}\left[\mathbf{K}\left(\mathbf{R}\widehat{\mathbf{f}}_i + \mathbf{t}\right)\right]_z &= \left[\mathbf{K}\left(\mathbf{R}\widehat{\mathbf{f}}_i + \mathbf{t}\right)\right]_x. \end{aligned}$$

where $[\mathbf{v}]_x$ denotes the x component of the vector $\mathbf{v}$. Then replacing $\widehat{\mathbf{f}}_i$ yields

$$\begin{aligned} f_{i,x}^{\mathrm{2D}}\left[\mathbf{K}\mathbf{R}\mathbf{M}_{2,i}{\mathbf{U}_{4,i}^{(2)}}^{\mathrm{T}}\mathbf{p}_2 + \mathbf{K}\mathbf{R}\bar{\mathbf{f}}_i + \mathbf{K}\mathbf{t}\right]_z &= \\ \left[\mathbf{K}\mathbf{R}\mathbf{M}_{2,i}{\mathbf{U}_{4,i}^{(2)}}^{\mathrm{T}}\mathbf{p}_2 + \mathbf{K}\mathbf{R}\bar{\mathbf{f}}_i + \mathbf{K}\mathbf{t}\right]_x & \end{aligned} \tag{39}$$

$$\begin{aligned} \Leftrightarrow \left(\left[\mathbf{K}\mathbf{R}\mathbf{M}_{2,i}{\mathbf{U}_{4,i}^{(2)}}^{\mathrm{T}}\right]_x - f_{i,x}^{\mathrm{2D}}\left[\mathbf{K}\mathbf{R}\mathbf{M}_{2,i}{\mathbf{U}_{4,i}^{(2)}}^{\mathrm{T}}\right]_z\right)\mathbf{p}_2 &= \\ f_{i,x}^{\mathrm{2D}}\left[\mathbf{K}\mathbf{R}\bar{\mathbf{f}}_i + \mathbf{K}\mathbf{t}\right]_z - \left[\mathbf{K}\mathbf{R}\bar{\mathbf{f}}_i + \mathbf{K}\mathbf{t}\right]_x. & \end{aligned} \tag{40}$$

Stacking the x- and y-components leads to

$$\begin{aligned} \begin{pmatrix} \left[\mathbf{K}\mathbf{R}\mathbf{M}_{2,i}{\mathbf{U}_{4,i}^{(2)}}^{\mathrm{T}}\right]_x - f_{i,x}^{\mathrm{2D}}\left[\mathbf{K}\mathbf{R}\mathbf{M}_{2,i}{\mathbf{U}_{4,i}^{(2)}}^{\mathrm{T}}\right]_z \\ \left[\mathbf{K}\mathbf{R}\mathbf{M}_{2,i}{\mathbf{U}_{4,i}^{(2)}}^{\mathrm{T}}\right]_y - f_{i,y}^{\mathrm{2D}}\left[\mathbf{K}\mathbf{R}\mathbf{M}_{2,i}{\mathbf{U}_{4,i}^{(2)}}^{\mathrm{T}}\right]_z \end{pmatrix}\mathbf{p}_2 &= \\ \begin{pmatrix} f_{i,x}^{\mathrm{2D}}\left[\mathbf{K}\mathbf{R}\bar{\mathbf{f}}_i + \mathbf{K}\mathbf{t}\right]_z - \left[\mathbf{K}\mathbf{R}\bar{\mathbf{f}}_i + \mathbf{K}\mathbf{t}\right]_x \\ f_{i,y}^{\mathrm{2D}}\left[\mathbf{K}\mathbf{R}\bar{\mathbf{f}}_i + \mathbf{K}\mathbf{t}\right]_z - \left[\mathbf{K}\mathbf{R}\bar{\mathbf{f}}_i + \mathbf{K}\mathbf{t}\right]_y \end{pmatrix}. & \end{aligned} \tag{41}$$

This equation system is extended to $2n$ rows by concatenating the two dimensions for each of the n corresponding points of one shape. Furthermore, one person parameter vector $\mathbf{p}_2$ can be estimated for multiple input shapes by stacking the points accordingly. Please note that the camera parameters $\mathbf{K}$, $\mathbf{R}$, $\mathbf{t}$ differ among shapes, but not among points of the same shape.

Similarly, for the expression parameter vector $\mathbf{p}_{34}$, we obtain

$$\begin{aligned} \begin{pmatrix} \left[\mathbf{K}\mathbf{R}\mathbf{M}_{34,i}{\mathbf{U}_{4,i}^{(4)}}^{\mathrm{T}}\right]_x - f_{i,x}^{\mathrm{2D}}\left[\mathbf{K}\mathbf{R}\mathbf{M}_{34,i}{\mathbf{U}_{4,i}^{(4)}}^{\mathrm{T}}\right]_z \\ \left[\mathbf{K}\mathbf{R}\mathbf{M}_{34,i}{\mathbf{U}_{4,i}^{(4)}}^{\mathrm{T}}\right]_y - f_{i,y}^{\mathrm{2D}}\left[\mathbf{K}\mathbf{R}\mathbf{M}_{34,i}{\mathbf{U}_{4,i}^{(4)}}^{\mathrm{T}}\right]_z \end{pmatrix}\mathbf{p}_{34} &= \\ \begin{pmatrix} f_{i,x}^{\mathrm{2D}}\left[\mathbf{K}\mathbf{R}\bar{\mathbf{f}}_i + \mathbf{K}\mathbf{t}\right]_z - \left[\mathbf{K}\mathbf{R}\bar{\mathbf{f}}_i + \mathbf{K}\mathbf{t}\right]_x \\ f_{i,y}^{\mathrm{2D}}\left[\mathbf{K}\mathbf{R}\bar{\mathbf{f}}_i + \mathbf{K}\mathbf{t}\right]_z - \left[\mathbf{K}\mathbf{R}\bar{\mathbf{f}}_i + \mathbf{K}\mathbf{t}\right]_y \end{pmatrix}. & \end{aligned} \tag{42}$$

In summary, when using a nonlinear camera model, the update equations for expression an person parameters are linear, corresponding to those in [2]. In the following section, we add constraints introduced above that leads to similar optimisation scheme for 2D input shapes as we proposed for the 3D.

4.3 Camera and Model Parameter Estimation for 2D Face Landmarks

4.3.1 Estimation of Model Parameters

Let us denote the linear equation (41) by $\mathbf{A}_2\mathbf{p}_2 = \mathbf{a}_2$. We seek to minimise the regularised energy functional

$$\begin{aligned} E_2^{\mathrm{p}}(\mathbf{p}_2) &= \frac{1}{2}\|\mathbf{A}_2\mathbf{p}_2 - \mathbf{a}_2\|_2^2 + \frac{\lambda_2}{2}\|{\mathbf{U}_4^{(2)}}^{\mathrm{T}}\mathbf{p}_2\|_2^2 + C' \\ &= \frac{1}{2}\mathbf{p}_2^{\mathrm{T}}\mathbf{Q}_2^{\mathrm{p}}\mathbf{p}_2 + {\mathbf{b}_2^{\mathrm{p}}}^{\mathrm{T}}\mathbf{p}_2 + C', \end{aligned} \tag{43}$$

where $\mathbf{Q}_2^{\mathrm{p}} := \mathbf{A}_2^{\mathrm{T}}\mathbf{A}_2 + \lambda_2\mathbf{U}_4^{(2)}{\mathbf{U}_4^{(2)}}^{\mathrm{T}}$ and $\mathbf{b}_2^{\mathrm{p}} := -\mathbf{A}_2^{\mathrm{T}}\mathbf{a}_2$.

In analogy to (32), by using the convex combination and neighbour constraints and by denoting $\mathbf{q}_2 = {\mathbf{P}_2^i}^{\mathrm{T}}\mathbf{p}_2$, we the minimisation problem takes the form

$$\min_i \min_{\mathbf{q}_2} \frac{1}{2}\mathbf{q}_2^{\mathrm{T}}\mathbf{Q}_2^{\mathrm{p},i}\mathbf{q}_2 + {\mathbf{b}_2^{\mathrm{p},i}}^{\mathrm{T}}\mathbf{q}_2, \tag{44}$$

subject to $\mathbf{q}_2 \geqslant \mathbf{0}$, $\mathbf{q}_2^{\mathrm{T}}\mathbf{1}_{\alpha_{\mathrm{P}}} = 1$, where $\mathbf{Q}_2^{\mathrm{p},i} = {\mathbf{P}_2^i}^{\mathrm{T}}\mathbf{Q}_2^{\mathrm{p}}\mathbf{P}_2^i$ and $\mathbf{b}_2^{\mathrm{p},i} = {\mathbf{P}_2^i}^{\mathrm{T}}\mathbf{b}_2^{\mathrm{p}}$.

Similarly, for $\mathbf{p}_{34}$ the minimisation problem is

$$\min_i \min_{\mathbf{q}_{34}} \frac{1}{2}\mathbf{q}_{34}^{\mathrm{T}}\mathbf{Q}_{34}^{\mathrm{p},i}\mathbf{q}_{34} + {\mathbf{b}_2^{\mathrm{p},i}}^{\mathrm{T}}\mathbf{q}_{34}, \tag{45}$$

Algorithm 2 3D Reconstruction and Camera Parameter Estimation from sparse 2D Landmarks

```
Input: $m$ 2D face landmark sets $\mathbf{f}_k^{2D}$, $k = 1, \dots, m$
• Initialisation:
    – Initialise $m$ cameras by DLT using mean face
    – Initialise $\widehat{\mathbf{p}}_{3,k}$ (or $\widehat{\mathbf{p}}_{34,k}$) with mean expression $\forall k$
• Repeat until convergence:
    – Model Parameter Estimation
      repeat until convergence:
        ∗ Given $\widehat{\mathbf{p}}_{34,k}$ and camera parameters, estimate person parameter vector $\widehat{\mathbf{p}}_2$, using Eq. (44)
        ∗ Given $\widehat{\mathbf{p}}_2$ and camera parameters, estimate expression parameter vectors $\widehat{\mathbf{p}}_{34,k}$, using Eq. (45)
    – Camera Parameter Estimation
        ∗ Given $\widehat{\mathbf{p}}_2$, $\widehat{\mathbf{p}}_{34,k}$, compute 3D shapes $\widehat{\mathbf{f}}_k^{3D}$
        ∗ Estimate camera parameters by $\widehat{\mathbf{f}}_k^{3D}$ and $\mathbf{f}_k^{2D}$ minimising Eq. (46).
Output: $\widehat{\mathbf{p}}_2$, $\widehat{\mathbf{p}}_{34,k}$, camera parameters, $\widehat{\mathbf{f}}_k^{2D}$, $\widehat{\mathbf{f}}_k^{3D}$
```

subject to $\mathbf{q}_{34} \geqslant \mathbf{0}$, $\mathbf{p}_4^{\mathrm{T}}\mathbf{1}_{\alpha_{\mathrm{E}}} = 1$, where $\mathbf{b}_{34}^{\mathrm{p},i} = -{\mathbf{P}_{34}^{i}}^{\mathrm{T}}\mathbf{A}_{34}^{\mathrm{T}}\mathbf{a}_{34}$ and $\mathbf{Q}_{34}^{\mathrm{p},i} = {\mathbf{P}_{34}^{i}}^{\mathrm{T}}\left(\mathbf{A}_{34}^{\mathrm{T}}\mathbf{A}_{34} + \lambda_{34}\mathbf{U}_4^{(4)}{\mathbf{U}_4^{(4)}}^{\mathrm{T}}\right)\mathbf{P}_{34}^{i}$. In effect, the same solver for the estimation of the model parameters for 3D and 2D input can be used, including the automatic determination of the weights λ_2 and λ_{34}.

4.3.2 *Camera Parameter Estimation*

Previously we assumed the camera parameters of all input shapes are given to approximate a 2D face shape $\mathbf{f}^{2\mathrm{D}}$ by a projected 3D model face shape $\widehat{\mathbf{f}}^{3\mathrm{D}}$ and then estimated the model parameters. Here we assume the model parameters are known, therefore the 3D face model shape $\widehat{\mathbf{f}}^{3\mathrm{D}}$ is given and projected onto the image plane yielding $\widehat{\mathbf{f}}^{2\mathrm{D}}$. The error between the original 2D landmarks $\mathbf{f}^{2\mathrm{D}}$ and the estimated 2D shape $\widehat{\mathbf{f}}^{2\mathrm{D}}$ is defined as

$$\epsilon_{\text{cam}} = \|\widehat{\mathbf{f}}^{2\mathrm{D}} - \mathbf{f}^{2\mathrm{D}}\|_2^2. \tag{46}$$

Given sparse correspondences between the 3D model points and the 2D landmarks, the parameters of the projective camera can be estimated by minimising Eq. (46) using a DLT (direct linear transform), see [35] for details. Please note that the global alignment is included in the camera parameter estimation procedure. The camera and model parameters can be estimated in an alternating scheme as described in Alg. 2, with the same stopping criteria as in Alg. 1.

5 Experiments

To build the model we used the Binghamton BU3DFE database [26] which consists of 2500 face scans of 100 persons performing 6 emotions (anger, disgust, fear, happiness, sadness, surprise) in 4 levels with increasing expression strength, and the neutral expression. For each shape 83 manually labelled landmark points are provided and used to build a sparse face model.

Dense correspondences among all face scans are computed by an adapted version of the ECPD [36] employing 3D landmarks to guide the registration. Thereafter a sparse or a dense multilinear face model can be estimated by Eq. (1), either based on 83 facial features points or the 7308 corresponding points.

5.1 Validation of the Apathy Vertex

In Section 2 the apathetic expression was identified as a natural origin, where each of the six prototypical emotions originate from. Here we show that it is not an artefact of the database or the factorisation approach.

5.1.1 *Apathy Mode and Expression Space of ADFES*

To confirm our hypothesis that this specific relaxed facial expression can be retrieved from posed facial expression databases other than BU3DFE, we choose a database with similar properties. The Amsterdam Dynamic Facial Expressions Set (ADFES) [27] contains image sequences of 22 persons, which starting in neutral thenchanging to full emotion (apex), varying in length. The emotions include the six basic emotions (anger, disgust, fear, joy, sadness, and surprise), which are the same as in the BU3DFE database, and neutral. We used the OpenFace [37] framework to detect $N = 68$ 2D landmarks for each frame. To create a data tensor from the ADFES database, all sequences of the six prototypical emotions and the neutral sequences were extracted. Then:

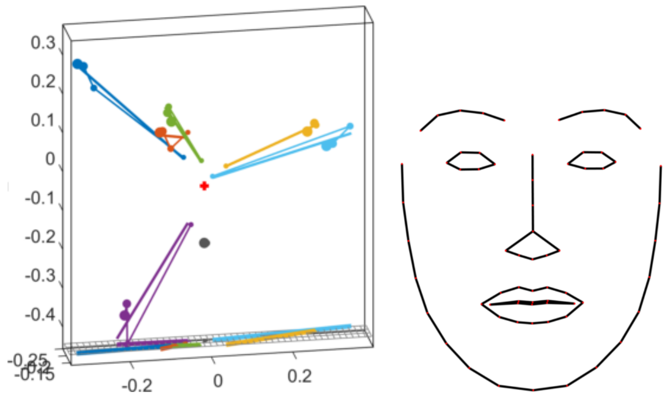
1) Globally align shapes, such that the top of the nose located at the origin.
2) From each sequence sample 4 frames, equidistantly.
3) The shapes are sorted into a 3D data tensor $\mathcal{T}_0 \in \mathbb{R}^{3N \times P \times EF}$, with $N = 68$, $P = 22$, $E = 6$, $F = 4$.
4) The mean shape is subtracted from $\mathcal{T} = \mathcal{T}_0 - \overline{\mathcal{T}}$, where $\overline{\mathcal{T}} \in \mathbb{R}^{3N \times P \times EF}$ contains the mean shape $\bar{\mathbf{f}}$, repeated to suit the size of the original tensor.
5) The expression space $\mathbf{U}^{(3)}$ is obtained by Eq. (1).
6) The apathy mode is estimated using $\mathbf{U}^{(3)}$.

The expression space $\mathbf{U}^{(3)}$ is depicted in Fig. 5(a), where the apathy mode is the red cross. The colours are analogue to Fig. 1. The expression space for ADFES is planar, star-shaped and contains linear trajectories for each emotion, just like it is for the BU3DFE as shown in Fig. 1. Fig. 5(b) displays the synthesised apathetic facial expression of the mean person for the AFDES, which is a relaxed facial expression with closed mouth. Based on these findings, we conclude that the previously discovered *apathy mode* is nor a result of overfitting, nor is it a property limited to one dataset.

5.1.2 *Justification Of Apathy Vertex in Face Shape Space*

An important question is whether the apparent intersection of the emotion trajectories at the point of apathy as indicated by Fig. 1 is an effect of the higher-order tensor factorisation or a bias of the BU3DFE dataset. We thus compare the $3N$-dimensional shape vectors of the *neutral* shapes with those closest to the intersections of the emotion trajectories both for BU3DFE and ADFES datasets. Since we cannot expect that $E = 6$ low-dimensional affine subspaces intersect in a single point in a high-dimensional space, we locate the point closest to all of the emotion trajectories, as in Sec. 2.1.1.

Let $\mathbf{f}_{e,p}^{k}$ denote the $3N$-dimensional shape vector of the kth out of $k = 1, \dots, 4$ expression levels of emotion e and person p. Denote by $\mathbf{v}_{e,p}^{l}$ with $l = 2, \dots, 4$ the difference



(a) expression space (b) apathetic expression

Figure 5. (a) Expression space for the ADFES, analogously to BU3DFE. (Colours as in Fig 1.) (b) apathetic facial expression synthesised for the average person.

$\mathbf{f}^l_{e,p} - \mathbf{f}^1_{e,p}$. Letting $\mathbf{V}_e$ be the matrix consisting of the vectors $\mathbf{v}^l_{e_i,p}$ of all $p = 1, \ldots, 100$ persons, we fit a 1-dimensional subspace to each $\mathbf{V}_e$. Let $\mathbf{B}_e$ denote the basis of the eth subspace, and $\bar{\mathbf{f}}_e = \frac{1}{P}\sum_p \mathbf{f}^1_{e,p}$ the average of shapes $\mathbf{f}^1_{e,p}$ of all persons.

The closest point $\mathbf{x}$ to each of the affine subspaces with basis $\mathbf{B}_e$ and origin $\bar{\mathbf{f}}_e$ w.r.t. to the world coordinate origin can be determined by solving the joint optimisation problem

$$\min_{\mathbf{x}} \sum_{e=1}^{E} \left\| \mathbf{x} - \left(\mathbf{P}_{\mathbf{B}_e} \left(\mathbf{x} - \bar{\mathbf{f}}_e \right) + \bar{\mathbf{f}}_e \right) \right\|_2^2 \tag{47}$$

where $\mathbf{P}_{\mathbf{B}_e}$ indicates the orthogonal projector onto the space spanned by $\mathbf{B}_e$.

The shape minimising Eq. (47) has a root mean square distance (RMSE) of 1.45 to each of the PE emotion trajectories for BU3DFE, whereas the average neutral shape has a RMSE of 2.45. The difference is not negligible since the mean squared distance between shapes of the same emotion is 7.81. The RMSE of the point closest to the intersection is 0.046 for ADFES, while the neutral point has an RMSE of 0.126 and the mean square distance in ADFES is 0.267. This confirms that the average neutral shape is more distant from the optimal centre of all emotion trajectories. The estimated shapes $\mathbf{x}$ looks very similar to the apathetic shape (Fig. 2).

5.2 Neutral vs. Apathy-centered Face Model

The original3D face shapes given in $\mathcal{T}_{4,0} \in \mathbb{R}^{3N \times P \times S \times E}$ can be approximated by an apathy-centered or neutral-centered tensor model. Assuming $\overline{\mathcal{T}}_{4,\text{apathy}}$ defines the tensor which consists of the mean apathetic face, while $\overline{\mathcal{T}}_{4,\text{neutral}}$ contains the mean neutral face shape. Then the 4D factorisation can be computed on the two versions, analogue to Eq. (5). This means for the apathy centered model, the difference tensor $\mathcal{T}_{4,a} = \mathcal{T}_{4,0} - \overline{\mathcal{T}}_{4,\text{apathy}}$ is approximated by

$$\mathcal{T}_{4,a} \approx \hat{\mathcal{T}}_{4,a} = \mathcal{S}_{4,a} \times_1 \mathbf{U}^{(1)}_{4,a} \times_2 \mathbf{U}^{(2)}_{4,a} \times_3 \mathbf{U}^{(3)}_{4,a} \times_4 \mathbf{U}^{(4)}_{4,a}. \tag{48}$$

The original shapes are then approximated by

$$\mathcal{T}_{4,0} \approx \mathcal{T}_{4,0,a} := \hat{\mathcal{T}}_{4,a} + \overline{\mathcal{T}}_{4,\text{apathy}}. \tag{49}$$

Analogously the original shapes are approximated by the neutral centered model as $\mathcal{T}_{4,0} \approx \mathcal{T}_{4,0,n}$. We found that the tensor using the apathy as centre gives a lower euclidean distance to the original shapes, than using the factorisation based on the centre of neutral, i.e.

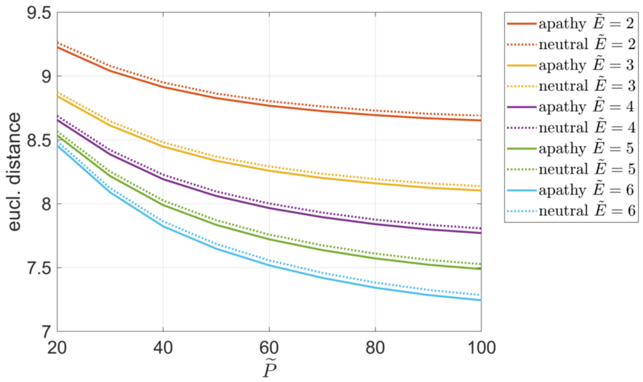


Figure 6. Change of the euclidean distance between true and estimated shapes, based on the apathy-centred (solid lines) and neutral-centred model (dashed lines) as in Eq. (50), for varying cropping factors. Two of the the four are fixed to $3\widetilde{N} = 250$ and $\widetilde{S} = 1$, while the cropping dimensions for person $\widetilde{P}$ and emotion $\widetilde{E}$ are varied.

$$\frac{1}{n}\|\mathcal{T}_{4,0,a} - \mathcal{T}_{4,0}\|_F < \frac{1}{n}\|\mathcal{T}_{4,0,n} - \mathcal{T}_{4,0}\|_F \tag{50}$$

where $\|.\|_F$ refers to the Frobenius norm, extended to tensors.[3] This relation was verified for the sparse and dense tensor, and various cropping factors. In Fig. 6 the decrease of the approximation error of Eq. (50) with respect to increased cropping factors for person and emotion dimension are visualised for $3\widetilde{N} = 250$ and $\widetilde{S} = 1$. The error obtained by the apathy-centred model (solid line) is always below the one of the neutral-centred model (dashed line). This means that the apathy-centred model retains more information in the first components than the neutral-centred model.

5.3 Residual Emotions of the Neutral Face

As several psychological studies suggest, there are six basic emotions that also form the natural basis for the expression analysis. In mathematical terms, the six emotions, centred at the point of apathy, form an affine basis where the apathetic expression is emotionless whereas all the other expressions are linear combinations of the basis emotions. The neutral face can be described in this basis, and it is expected to reflect *residual emotions* [21]. This is illustrated in Fig. 7 where we show the expression change when interpolating from angry to disgust via either neutral or apathetic expression. In the top row the face in the gray box is labelled as neutral in BU3DFE, although it looks rather happy. In contrast to that in the second row the trajectory passes through the apathetic, emotionless face, which does not change the overall emotion.

5.4 Person and Expression Transfer

In this section, we compare the different parameterisations of tensor models shown in Tab. 1, which includes the model *sub+* as an intermediate model between *sub* and *4D*. *sub* and *sub+* are both based on a mean-centred 3D data tensor

[3] Frobenius norm for tensors: $\|\mathcal{T}\|_F = \sqrt{\sum_{i,p,l,e} |t_{iple}|^2}$

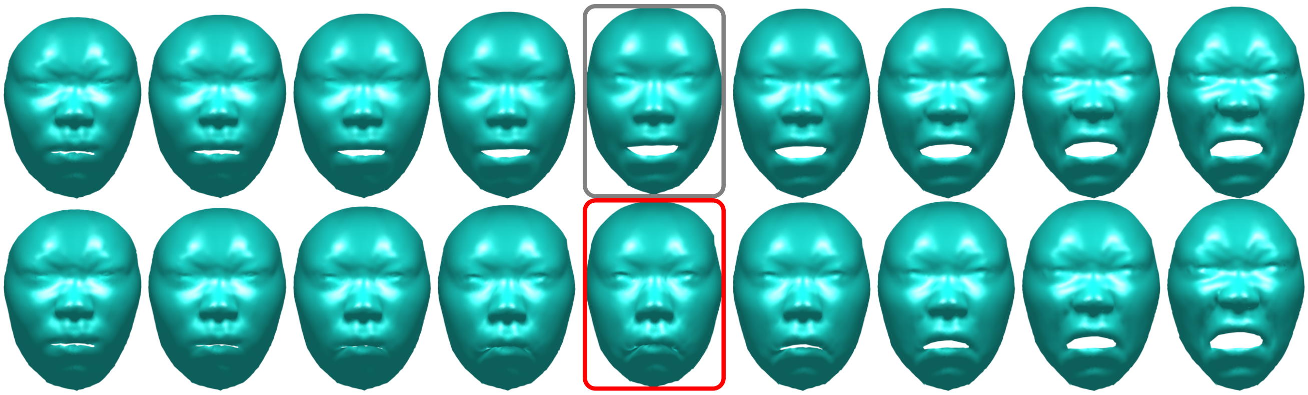
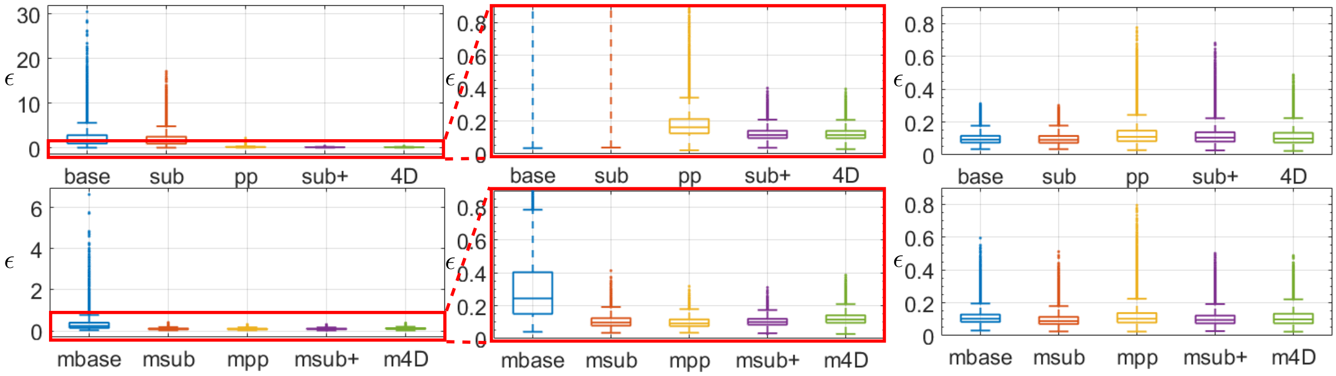


Figure 7. Synthesised expression trajectories, both starting in anger and ending in disgust. In the first line intersecting the neutral expression (grey box), whereas in the second row the face in the red box represents the synthesised apathetic facial expression.



(a) expression transfer (b) expression transfer (c) person transfer

Figure 8. Quantitative evaluations of the methods by Eq. (51) in (a), (b) expression transfer; and (c) person transfer. (first row) Person and expression parameter vectors person are estimated independently for each shape; (second row) a shared person parameter vector for varying expressions performed by a single person is estimated. It can be seen that, in expression transfer, the sparsity constraint of the models *sub+* and *4D* and the shared estimation of the person parameter vector yields the best, more robust performance than the reference methods. In person transfer, there are no significant differences between the methods, except *pp* seems less robust yielding more outliers; the sharing of the person parameter vector seem to improve *sub+* and *4D*. Since *sub+* and *4D* yield similar result in all the experiments, it is an empirical proof that the emotion strength subspace can be safely truncated into rank-one subspace. The abbreviations are defined in Tab. 1.

Table 1
Abbreviations for different tensor model parameterisations.

Label	Model Description
base	The baseline model defined in Eq. (2)
sub	The subspace aware model from [1] with parameters as in Eq. (3) and constraints from [1]
sub+	Same as *sub* augmented with the constraints of *4D*
pp	Projection pursuit model from [1]
4D	4D model defined by Eq. (9), (22), (32)-(33)
m (prefix)	Person parameter vector shared among shapes of the same person

employing 25 expression parameters, whereas *4D* is centred at apathy and employs only 6 expression parameters. The difference between *sub* and *sub+* is that the optimisation for *sub* is done by a direct linear equation system, while the base for *sub+* and *4D* are given by Eq. (32)-(33) with constraints for sparsity and positivity. The comparison is done by person and expression transfer by leave-one-out experiments. This means that for an unknown 3D face shape, we estimate person and expression parameters by Alg. 1 for the models, using the sparse model representation with $N = 83$ points. Then either the person or expression parameter vector is changed to known values to perform either person or expression transfer. If the person parameter vector was estimated reasonably, changing the expression parameter is expected to alter the expression only. Otherwise, the worst result would be a degeneration of the shape. The error between an estimated shape $\hat{\mathbf{f}}$ and the true shape $\mathbf{f}_{\text{true}}$ is defined by

$$\epsilon\left(\hat{\mathbf{f}}, \mathbf{f}_{\text{true}}\right) := \frac{\|\hat{\mathbf{f}} - \mathbf{f}_{\text{true}}\|_2}{\|\mathbf{f}_{\text{true}}\|_2}. \quad (51)$$

A 3D face shape based on the person and expression parameter vectors $\mathbf{w}_2, \mathbf{w}_3$ is denoted as $\mathbf{f}^{3\text{D}}(\mathbf{w}_2, \mathbf{w}_3)$. The experiment will be performed as described in Alg. 3. For each experiment three errors are computed: approximation, expression transfer and person transfer. For each model, the most versatile parameter setting is defined as the one which gives the minimum median error over the three errors. The experiment is repeated by leaving one level out instead of one person. Additionally since subsets of the excluded shapes stem from the same person, it is reasonable to demand one joint person parameter vector among them.

Algorithm 3 Leave-One-Person-Out Experiment

```
Input: Data tensor $\mathcal{T}_{3,0} \in \mathbb{R}^{3N \times P \times E_{tot}}$, $E_{tot} = S \cdot E$ without neutral expression
  1) Exclude person $p$ means excluding shapes $\mathbf{f}^{3\mathrm{D}}_{p,e}$, $e = 1, \ldots, E_{tot}$, gives the reduced data tensor $\mathcal{T}_p \in \mathbb{R}^{3N \times (P-1) \times E_{tot}}$.
  2) Re-estimate the HOSVD on the reduced data tensor
         $\mathcal{T}_{3,p} = \mathcal{S}_{3,p} \times_1 \mathbf{U}^{(1)}_{3,p} \times_2 \mathbf{U}^{(2)}_{3,p} \times_3 \mathbf{U}^{(3)}_{3,p}$   (52)
     If the model $4D$ is used, proceed with these two steps
     • Re-estimate the apathy-vertex from $\mathbf{U}^{(3)}_p$
     • Compute the $4D$ apathy-centred data tensor $\mathcal{T}_{4,p} \in \mathbb{R}^{3D \times (P-1) \times S \times E}$ and the HOSVD
         $\mathcal{T}_{4,p} = \mathcal{S}_{4,p} \times_1 \mathbf{U}^{(1)}_{4,p} \times_2 \mathbf{U}^{(2)}_{4,p} \times_3 \mathbf{U}^{(3)}_{4,p} \times_4 \mathbf{U}^{(4)}_{4,p}$   (53)
  3) Estimate model parameters for each excluded shape by Alg. 1 and obtain the approximated 3D faces $\mathbf{f}^{3\mathrm{D}}_{p,e}$.
  4) Expression transfer is performed by replacing the estimated expression parameter vector $\widehat{\mathbf{w}}_3$ (or $\widehat{\mathbf{w}}_{34}$) by its true value $\mathbf{w}_{34}$, while keeping the estimated person parameter vector $\widehat{\mathbf{w}}_2$ constant, the 3D face shape of person $p$ in expression $e$ is computed using $(\widehat{\mathbf{w}}_2, \mathbf{w}_{34})$.
  5) Person transfer is performed by replacing the estimated person parameter $\widehat{\mathbf{w}}_2$ by a known parameter vector of one of the remaining $P - 1$ persons, using the parameters $(\mathbf{w}_2, \widehat{\mathbf{w}}_{34})$ to compute the new shape.
  6) The transfer errors are computed by using Eq. (51), i.e.:
     for expression $\epsilon\left(\mathbf{f}^{3\mathrm{D}}(\widehat{\mathbf{w}}_2, \mathbf{w}_{34}), \mathbf{f}^{3\mathrm{D}}_{p,e}\right)$ and
     for person $\epsilon\left(\mathbf{f}^{3\mathrm{D}}(\mathbf{w}_2, \widehat{\mathbf{w}}_{34}), \mathbf{f}^{3\mathrm{D}}_{p,e}\right)$.
Output: Errors $\epsilon$.
```

Therefore the experiments are performed for all tensor models listed in Tab. 1, and either estimate the parameter vectors individually for each shape or constrain the same person parameter vector.

The first row in Fig 8 shows results based on person and expression parameter vectors estimated individually for each left-out-shape, whereas in the second row the person parameter vector is shared among varying expressions performed by the same person, hence the prefix *m*. Comparing the top and bottom row shown in Fig. 8(a)-(b) it can be seen that the expression transfer error decreases if the person parameter vector is estimated based on several shapes. However considering often only one shape per person is available, the first row is more relevant. It shows that the error decreases with each model variant, while the models *sub+* and *4D* lead to similar results and perform best. The models perform similarly for person transfer depicted in Fig. 8(c), which we assume to be a missing evidence in the underlying training data. Please note that the approximation error (not depicted) leads to similarly good results for all models. In general since *sub+* and *4D* yield similar result in all the experiments, it is an empirical proof that the emotion strength subspace can be safely truncated into a rank-one subspace. In conclusion, the latest proposed model *4D* is best, because it has fewer parameters than all the others but is still able to perform equally good or even better than the other models, hence is more robust. This is a consequence of the bias–variance dilemma. The results of model *sub+* and *4D* are based on $\alpha_P = 5$, $\alpha_E = 2$.

5.5 Influence of Parameters

To determine the influence of the parameters on the result, we perform experiments by Alg. 3 by leaving one person or level out with varying number of neighbours for emotion α_E and person α_P for our model *4D*. In Fig. 9 each square represents the mean value of the two settings, given the number of neighbours of emotion α_E on the y-axis, and for person α_P on the x-axis. The approximation error is lowest if the number of neighbours is largest, while the transfer error for person and expression is small if the number of neighbours is small. In conclusion the parameters should be selected depending on the desired application.

5.6 3D Reconstruction from 2D Landmarks

In this section, we reconstruct dense 3D face shapes from sparse 2D landmarks. In contrast to [38] presenting a 3D shape regression based on the FW database, we do not rely on extensive training of 60 images per unseen person. In [39], 3D reconstructions with more details are presented, but require user intervention and a second camera for an initial blendshape model fitting. Neither of them evaluate on a databases of 3D faces. In this section we use the estimation scheme presented in Sec. 4 using sparse 2D landmarks. The model abbreviations of Tab. 1 are as before.

5.6.1 3D Reconstructions of Bosphorus Database

The Bosphorus database [40] consists of images and 3D face scans of 105 individuals varying in facial expression. We choose the face scans annotated with the seven basic emotions. Because 178 of potential 735 datasets are missing, the total number is 557. For each image we detected landmarks using OpenFace [41], [42], which implies dlib [43].

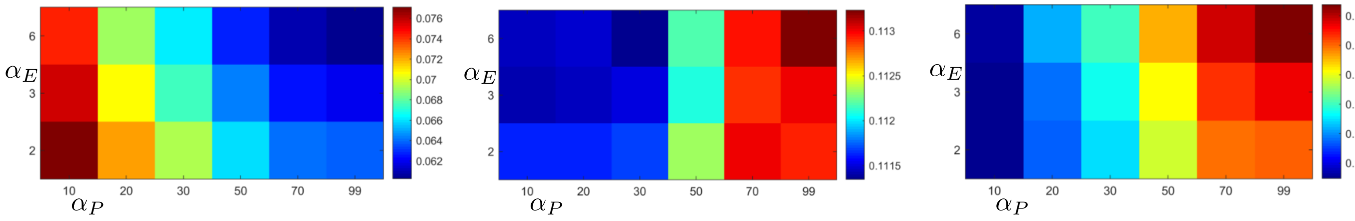
Hereafter different approaches are used to estimate dense 3D reconstructions from 2D. The *Surrey Face Model (SFM)* [44] is based on a 3D morphable model, i.e. PCA-based. We apply their code [45], which extends [44] by varying expressions and image edge information [46]. Additionally a neural network approach designed for detailed 3D reconstruction is used, referred to as *Sela* [47], code [48].

Different versions of the tensor face model based on the databases BU3DFE [26] and BU4DFE [49]. The BU4DFE is similar to the BU3DFE, but consists of sequences of 3D face scans, which we temporally aligned [50] to 10 samples per person conforming to 10 levels (expression intensities) from neutral to full emotion. In the following the models *base* and *pp* are excluded, because the base model is highly unstable in conjunction with a projective camera, whereas the results of *pp* are almost the same as for *sub*.

The points of the true 3D face scan are defined as $\mathbf{p}_l \in \mathbb{R}^3$, for the estimated as $\widehat{\mathbf{p}}_k$, and their point-wise correspondence is $I = \{(i,j) | \widehat{\mathbf{p}}_i \text{ corresponds to } \mathbf{p}_j\}$. The mean squared euclidean distance between them is

$$Q = \frac{1}{|I|} \sum_{(i,j) \in I} \|\widehat{\mathbf{p}}_i - \mathbf{p}_j\|_2^2, \tag{54}$$

where a prior rigid alignment of the 3D faces is assumed. Because 7 samples for each person are provided, we can



(a) Approximation $\epsilon\left(\mathbf{f}^{3D}(\hat{\mathbf{w}}_2, \hat{\mathbf{w}}_{34}), \mathbf{f}^{3D}_{p,e}\right)$ (b) Expression transfer $\epsilon\left(\mathbf{f}^{3D}(\hat{\mathbf{w}}_2, \mathbf{w}_{34}), \mathbf{f}^{3D}_{p,e}\right)$ (c) Person transfer $\epsilon\left(\mathbf{f}^{3D}(\mathbf{w}_2, \hat{\mathbf{w}}_{34}), \mathbf{f}^{3D}_{p,e}\right)$

Figure 9. Influence of number of neighbours of emotion α_E (y-axis) and person α_P (x-axis) on the different error measures, see Alg. 3.

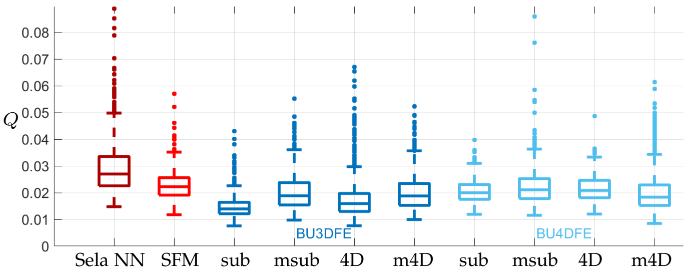


Figure 10. Boxplot of mean euclidean distances between true and estimated shapes, as in Eq. (54). The colours refer to different training databases and the labels on the x-axis refer to varying models.

demand the same person parameter vector must apply. In Fig. 10 databases are distinguished by colour, while the name of the model is the x-axis label, as in Tab. 1. The two reference models (red) perform worse than all variants of the tensor models. Comparing the tensor models *sub* and *4D* shows that both median estimates lie within the error bounds of the other that shows that there is no significant difference in their performance. Moreover, this is the numerical justification of the truncation of the emotion strength subspace into a one-dimensional subspace as no significant degradation of the result occurs. Selected qualitative examples are provided in Fig. 11. While the results of *Sela NN* are deformed, *SFM* leads to more stable results. Also as expected the expressiveness of our models decreases slightly for multiple inputs. The results of model *4D* are based on $\alpha_E = 2$, $\alpha_P = 5$.

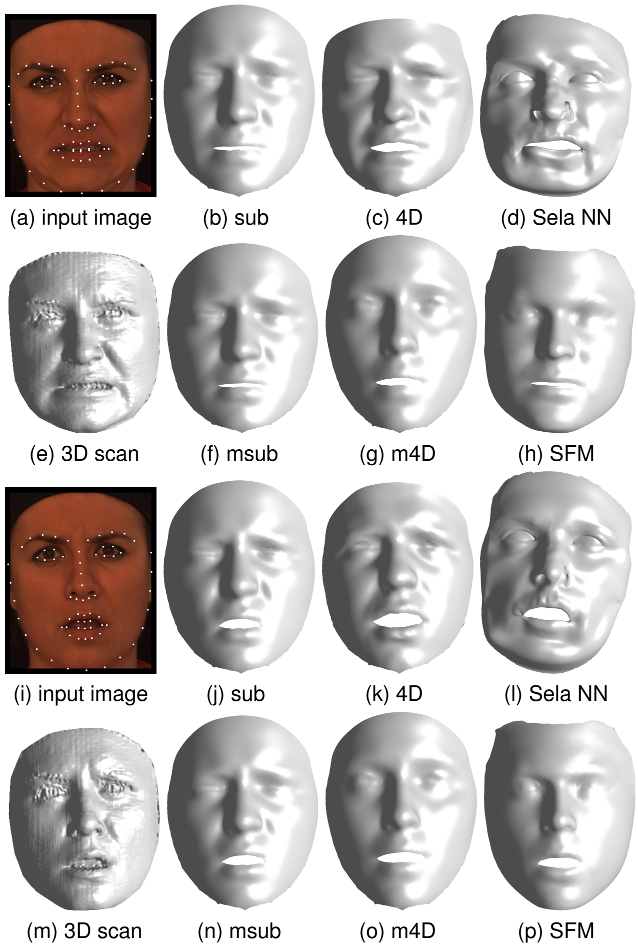

(a) input image (b) sub (c) 4D (d) Sela NN
(e) 3D scan (f) msub (g) m4D (h) SFM
(i) input image (j) sub (k) 4D (l) Sela NN
(m) 3D scan (n) msub (o) m4D (p) SFM

Figure 11. Results of 3D reconstructions from person 1. In (a),(i) original images in expressions disgust and fear with ground truth 3D scan in (e),(m). The remainder are 3D reconstructions based on the models as indicated. The proposed models: (b),(j) *sub*, (c),(k) *4D*, (f),(n) *msub*, (g),(o) *m4D*, and the reference models in (d),(l) Sela NN, and (d),(l) SFM.

5.6.2 3D Reconstructions in the Wild

The Bosphorus database was recorded in a highly controlled environment. The Florence database [51] contains data of 53 persons, with at least one 3D face scan and one to three video sequences in challenging environments of varying length. Because there is no 3D scan for each frame of the video, a comparison between estimated and true 3D faces is not possible. Therefore we give the 2D error, i.e. the mean pixelwise distance between the given landmarks and the corresponding projected 3D points of our models *sub* and *4D*. As before we estimated 3D reconstructions from landmarks (obtained by the OpenFace [42]) for 228 selected frames. We found that the model *sub* gives a mean euclidean pixel distance of 5.4, and 6.1 for *4D*. Considering that our model only employs a sparse subset of 46 landmarks with no additional prior knowledge, these results seem reasonable. Additionally we repeated the experiment on a subset of the AFLW [52] database offered by [25]. The mean euclidean distance between the six most stable landmarks (corners of eyes and mouth) and their corresponding estimated reprojected 3D points turned out 6.9 for the reference [25] and 3.6 for our *4D*, which confirms that our model is on-par with state-of-the-art methods.

To finalise the in the wild experiments we selected two more examples, and estimated 3D reconstructions just as in Sec. 5.6.1. (Additional examples are in the supplemental material.) The results shown in Fig. 12 show that the proposed model *4D* with few parameters utilising only 46 landmarks

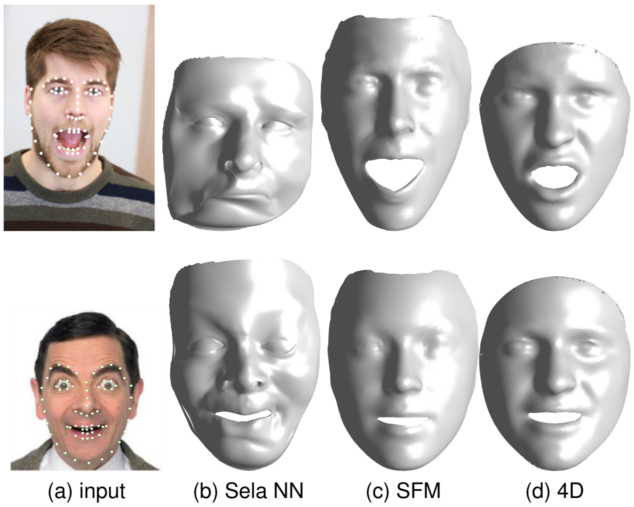


Figure 12. Dense 3D reconstruction examples based on input (a), and outcomes based on (b) Sela NN, (c) SFM, and (d) our 4D model.

as input gives satisfying 3D reconstructions on par with the state of the art methods. In the second row shown in Fig. 12 reveals some limitation of the training data, i.e. since an expression with smiling mouth and raised eyebrows is not part of the training data, it cannot be recovered. Therefore we conclude that the results reflect a limitation of the training data, not of the model parameterisation.

6 Conclusion

In this work, we proposed extensions to our tensor-based 3D face model [1], where we showed that the facial expression space prevails a star-shaped substructure for a 3D face database while an apathetic facial expression lies in the origin of this affine subspace. We employ this knowledge by constructing an apathy-centred expression space that yields a more compact 4D tensor model in contrast to the earlier model. The 4D model describes the expressions with 6 parameters opposed to the earlier 25 by folding the expression strengths into an additional mode of the tensor. Moreover, the expression strength mode of the tensor can be heavily truncated into a one-dimensional subspace that leads to the proposed compact model. A sparsity constraint on the subspace parameters for person and expression, controlled by one value each, compresses the model further, leading to a more robust version. For the optimisation procedure, we introduced an automatic way of computing regularisation parameters, thus avoiding manual user intervention and time-consuming parameter tuning. In addition, we proposed a 3D reconstruction method for faces and expression from sparse 2D landmarks by assuming projective camera without calibration information. Our experiments confirmed that transfer of person and expression can be performed better or equally well if compared to the previous models while the proposed model is more compact, i.e., fewer parameters are required. In the experiments, we also validated the existence of the star-shaped structure of the expression space by another database of 2D facial expressions. On the basis of this work, we conclude that the proposed tensor-based model is a compact accurate descriptor for faces and expressions, and hence a promising tool for various applications.

References

[1] S. Grasshof, H. Ackermann, S. Brandt, and J. Ostermann, "Apathy is the root of all expressions," in *Proc. FG*, 2017.

[2] S. Grasshof, H. Ackermann, J. Ostermann, and S. Brandt, "Projective structure from facial motion," in *Proc. MVA*, 2017.

[3] B. Allen, B. Curless, and Z. Popović, "The Space of Human Body Shapes: Reconstruction and Parameterization from Range Scans," in *ACM SIGGRAPH*, 2003, pp. 587–594.

[4] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis, "SCAPE: Shape Completion and Animation of People," in *ACM SIGGRAPH*, vol. 24, no. 3, 2005, pp. 408–416.

[5] N. Hasler, C. Stoll, M. Sunkeln, B. Rosenhahn, and H.-P. Seidel, "A statistical model of human pose and body shape." *Computer Graphics Forum (Proc. Eurographics)*, vol. 28, pp. 337–346, 2009.

[6] N. Hasler, H. Ackermann, B. Rosenhahn, T. Thormählen, and H.-P. Seidel, "Multilinear pose and body shape estimation of dressed subjects from image sets," in *Proc. CVPR*, Jun. 2010, pp. 1823–1830.

[7] B. Wandt, H. Ackermann, and B. Rosenhahn, "3d reconstruction of human motion from monocular image sequences," *IEEE T Pattern Anal*, vol. 38, no. 8, pp. 1505–1516, 2016.

[8] ——, "A kinematic chain space for monocular motion capture," in *Proc. ECCV WS*, Sep. 2018.

[9] M. Turk and A. Pentland, "Eigenfaces for Recognition," *J Cognitive Neurosci*, vol. 3, no. 1, pp. 71–86, Jan. 1991.

[10] M. A. O. Vasilescu and D. Terzopoulos, "Multilinear Analysis of Image Ensembles: TensorFaces," in *Proc. ECCV*, 2002, no. 2350, pp. 447–460.

[11] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: a factorization method," *Int J Comput Vision*, vol. 9, no. 2, pp. 137–154, Nov. 1992.

[12] C. Bregler, A. Hertzmann, and H. Biermann, "Recovering non-rigid 3d shape from image streams," in *Proc. CVPR*, vol. 2, 2000, pp. 690–696 vol.2.

[13] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Proc. SIGGRAPH*, 1999, pp. 187–194.

[14] B. Amberg, R. Knothe, and T. Vetter, "Expression invariant 3d face recognition with a morphable model," in *Proc. FG*, Sep. 2008.

[15] B. Chu, S. Romdhani, and L. Chen, "3d-Aided Face Recognition Robust to Expression and Pose Variations," in *Proc. CVPR*, 2014, pp. 1907–1914.

[16] P. Garrido, M. Zollhöfer, D. Casas, L. Valgaerts, K. Varanasi, P. Pérez, and C. Theobalt, "Reconstruction of Personalized 3d Face Rigs from Monocular Video," *ACM T Graphic (SIGGRAPH)*, vol. 35, no. 3, pp. 28:1–28:15, May 2016.

[17] J. Thies, M. Zollhöfer, M. Niessner, L. Valgaerts, M. Stamminger, and C. Theobalt, "Real-time expression transfer for facial reenactment," *ACM T Graphic (SIGGRAPH)*, vol. 34, no. 6, Oct. 2015.

[18] D. Vlasic, M. Brand, H. Pfister, and J. Popović, "Face Transfer with Multilinear Models," in *ACM SIGGRAPH*, 2005, pp. 426–433.

[19] A. Brunton, T. Bolkart, and S. Wuhrer, "Multilinear Wavelets: A Statistical Shape Space for Human Faces," in *Proc. ECCV*, Jan. 2014, pp. 297–312.

[20] A. Jain, T. Thormählen, H.-P. Seidel, and C. Theobalt, "MovieReshape: Tracking and reshaping of humans in videos," in *ACM T Graphic (SIGGRAPH Asia)*, 2010.

[21] E. Lee, J. Kang, I. H. Park, J.-J. Kim, and S. An, "Is a neutral face really evaluated as being emotionally neutral?" *Psychiatry research*, vol. 157, pp. 77–85, 02 2008.

[22] D. Joo, D. Kim, and J. Kim, "Generating a fusion image: One's identity and another's shape," in *Proc. CVPR*, June 2018.

[23] F.-J. Chang, A. T. Tran, T. Hassner, I. Masi, R. Nevatia, and G. Medioni, "Deep, landmark-free fame: Face alignment, modeling, and expression estimation," *Int J Comput Vision*, vol. 127, no. 6, pp. 930–956, Jun 2019.

[24] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *Proc. AVSS*, 2009.

[25] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3d solution," in *Proc. CVPR*, 2016, pp. 146–155.

[26] L. Yin, X. Wei, Y. Sun, J. Wang, and M. Rosato, "A 3d facial expression database for facial behavior research," in *Proc. FG*, Apr. 2006, pp. 211–216.

[27] J. van der Schalk, S. T. Hawk, A. H. Fischer, and B. Doosje, "Moving faces, looking places: Validation of the Amsterdam Dynamic Facial Expression Set (ADFES)," *Emotion*, vol. 11, no. 4, pp. 907–920, 2011.

[28] L. De Lathauwer and B. De Moor, "A multi-linear singular value decomposition," *Society for Industrial and Applied Mathematics*, vol. 21, pp. 1253–1278, 03 2000.

[29] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "FaceWarehouse: A 3d Facial Expression Database for Visual Computing," *IEEE T Vis Comput Gr*, vol. 20, no. 3, pp. 413–425, Mar. 2014.

[30] K. H. Jensen, F. J. Sigworth, and S. S. Brandt, "Removal of vesicle structures from transmission electron microscope images," *IEEE T Image Process*, vol. 25, no. 2, pp. 540–552, 2016.

[31] K. Kanatani, Y. Sugaya, and H. Ackermann, "Uncalibrated Factorization Using a Variable Symmetric Affine Camera," in *Proc. ECCV*, May 2006, pp. 147–158.

[32] S. Brandt, "Closed-form solutions for affine reconstruction under missing data," in *Proc. SMVP in conjunction to ECCV*, 2002, pp. 109–114.

[33] H. Ackermann and K. Kanatani, "Iterative low complexity factorization for projective reconstruction," in *Proc. RobVis*, 2008, pp. 153–164.

[34] H. Ackermann and B. Rosenhahn, "Projective Reconstruction from Incomplete Trajectories by Global and Local Constraints," in *Proc. CVMP*, 2011, pp. 77–86.

[35] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge: Cambridge University Press, 2004.

[36] V. Golyanik, B. Taetz, G. Reis, and D. Stricker, "Extended coherent point drift algorithm with correspondence priors and optimal subsampling," in *Proc. WACV*, 2016, pp. 1–9.

[37] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," CMU-CS-16-118, CMU School of Computer Science, Tech. Rep., 2016.

[38] C. Cao, Y. Weng, S. Lin, and K. Zhou, "3d shape regression for real-time facial animation," *ACM Trans. Graph.*, vol. 32, no. 4, pp. 41:1–41:10, Jul. 2013.

[39] P. Garrido, L. Valgaert, C. Wu, and C. Theobalt, "Reconstructing detailed dynamic face geometry from monocular video," *ACM Trans. Graph.*, vol. 32, no. 6, pp. 158:1–158:10, Nov. 2013.

[40] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus Database for 3d Face Analysis," in *Biometrics and Identity Management*, 2008, pp. 47–56.

[41] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. CVPR*, Jun. 2014, pp. 1867–1874.

[42] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," CMU-CS-16-118, CMU School of Computer Science, Tech. Rep., 2016.

[43] D. E. King, "Dlib-ml: A machine learning toolkit," *J Mach Learn Res*, vol. 10, pp. 1755–1758, 2009.

[44] P. Huber, G. Hu, R. Tena, P. Mortazavian, P. Koppen, W. J. Christmas, M. Ratsch, and J. Kittler, "A Multiresolution 3d Morphable Face Model and Fitting Framework," in *Proc. VISIGRAPP*, Feb. 2016.

[45] P. Huber, "Surrey face model," https://github.com/patrikhuber/eos, 2018.

[46] A. Bas, W. A. P. Smith, T. Bolkart, and S. Wuhrer, "Fitting a 3D Morphable Model to Edges: A Comparison Between Hard and Soft Correspondences," in *ACCV WS on Facial Inf*, Nov. 2016.

[47] M. Sela, E. Richardson, and R. Kimmel, "Unrestricted facial geometry reconstruction using image-to-image translation," in *Proc. ICCV*, 2017, pp. 1585–1594.

[48] ——, "Unrestricted facial geometry reconstruction using image-to-image translation," *github*, 2017. [Online]. Available: https://github.com/matansel/pix2vertex

[49] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, "A high-resolution 3d dynamic facial expression database," in *Proc. FG*, Sep. 2008.

[50] M. Awiszus, S. Graßhof, F. Kuhnke, and J. Ostermann, "Unsupervised features for facial expression intensity estimation over time," in *Proc. CVPR WS*, Jun. 2018.

[51] A. D. Bagdanov, A. Del Bimbo, and I. Masi, "The florence 2d/3d hybrid face dataset," in *Proceedings of the 2011 Joint ACM Workshop on Human Gesture and Behavior Understanding*, ser. J-HGBU '11. New York, NY, USA: ACM, 2011, p. 79–80. [Online]. Available: http://doi.acm.org/10.1145/2072572.2072597

[52] P. M. R. Martin Koestinger, Paul Wohlhart and H. Bischof, "Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization," in *Proc. First IEEE Int. Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.

Stella Graßhof studied Computational Life Science at the University of Lübeck, where she received her M.Sc. in 2012. Her thesis focused on nonrigid image registration of vector valued data. Since 2012 she is working as a PhD student at the Institut für Informationsverarbeitung (TNT) at Leibniz Universität Hannover. Her research interests are face analysis and modeling, 3D reconstruction, facial animation, image registration, computer vision, and machine learning. Since 08/2019 she is working as a PostDoc at the IT University of Copenhagen.

Hanno Ackermann studied Computer Engineering at the University of Mannheim. He received his masters degree (Dipl.-Inf.) in 2003. From 10/2004 until 3/2008 he did his Phd at the University of Okayama, Japan. From 5/2008 until 9/2008 he worked as PostDoc at the Max-Planck-Institute for Computer Science in Saarbruecken, Germany. Since 10/2008 he is a member of the group of Prof. Rosenhahn at Leibniz University Hannover. He is currently funded by the German Research Foundation (DFG RO 2497/12-2). He is interested in theoretical and practical aspects of supervised and unsupervised learning, segmentation and clustering of data, model and pattern detection as well as model fitting under incomplete and corrupt data.

Sami Sebastian Brandt got his doctoral degree in Helsinki University of Technology, Finland, in 2002 and habilitated on the geometric branch of computer vision in University of Oulu, Finland, in 2007. After the doctoral degree he worked for one year as a research scientist in Instrumentarium Corporation Imaging Division, Finland, a couple of years in Helsinki University of Technology, University of Oulu, Finland, Malmö University, Sweden, Nordic Bioscience Imaging/Synarc Imaging Technologies, and 3Shape in Denmark. He currently has a faculty position as associate professor in the Department of Computer Science, IT University of Copenhagen, and additionally at the Image Group in University of Copenhagen. His research interests include applied mathematics, statistical inverse problems, Bayes methods, electron tomography, single particle reconstruction, geometric computer vision, image analysis, and machine learning.

Jörn Ostermann studied Electrical Engineering and Communications Engineering at the University of Hannover and Imperial College London. He received Dipl.-Ing. and Dr.-Ing. from the University of Hannover in 1988 and 1994, respectively. In 1994, he joined AT&T Bell Labs. From 1996 to 2003 he was with AT&T Labs – Research. Since 2003 he is Full Professor and Head of the Institut für Informationsverarbeitung at the Leibniz Universität Hannover, Germany. Since 2008, Jörn is the Chair of the Requirements Group of MPEG (ISO/IEC JTC1 SC29 WG11). Jörn received several international awards and is a Fellow of the IEEE. His current research interests are video coding and streaming, computer vision, 3D modeling, face animation, and computer–human interfaces.