Disentangling the Modes of Variation in Unlabelled Data

Mengjiao Wang, Yannis Panagakis,Patrick Snape, and Stefanos Zafeiriou, *Member, IEEE*

Abstract—Statistical methods are of paramount importance in discovering the modes of variation in visual data. The Principal Component Analysis (PCA) is probably the most prominent method for extracting a single mode of variation in the data. However, in practice, several factors contribute to the appearance of visual objects including pose, illumination, and deformation, to mention a few.To extract these modes of variations from visual data, several supervised methods, such as the TensorFaces relying on multilinear (tensor) decomposition have been developed. The main drawbacks of such methods is that they require both labels regarding the modes of variations and the same number of samples under all modes of variations (e.g., the same face under different expressions, poses etc.). Therefore, their applicability is limited to well-organised data, usually captured in well-controlled conditions. In this paper, we propose a novel general multilinear matrix decomposition method that discovers the multilinear structure of possibly incomplete sets of visual data in unsupervised setting (i.e., without the presence of labels). We also propose extensions of the method with sparsity and low-rank constraints in order to handle noisy data, captured in unconstrained conditions. Besides that, a graph-regularised variant of the method is also developed in order to exploit available geometric or label information for some modes of variations. We demonstrate the applicability of the proposed method in several computer vision tasks, including Shape from Shading (SfS) (in the wild and with occlusion removal), expression transfer, and estimation of surface normals from images captured in the wild.

Index Terms—Unsupervised multilinear decomposition, tensor decomposition, shape from shading, expression transfer

1 Introduction

Statistical methods that explain variability among observed measurements (data) in terms of a potentially lower number of unobserved, latent, variables are cornerstones in data analysis, image and signal processing, and computer vision.

Factor analysis [1] and the closely related Principal Component Analysis (PCA) [2] and Singular Value Decomposition (SVD) are probably the most popular statistical methods to find a single mode of variation that explains the data. Nevertheless, most forms of (visual) data have many different and possibly independent, modes of variations and hence methods such as the PCA are not able to identify them. Consider, for example, a population of faces with differing identities and expressions observed under different views (poses) where the appearance of each face is a result of some multifactor confluence due to identity, expression, and pose variation. In order to disentangle multiple but independent modes of variations, several multilinear (tensor) decompositions have been employed [3], [4], [5], [6], [7]. For instance, the High Order SVD (HOSVD) [4] is able to identify different modes of variation for identities, expressions, and poses per pixel, from a population of faces, by decomposing a carefully designed data tensor. This method is known as TensorFaces [8].

The main limitation of the above multilinear decompositions in disentangling multiple modes of variation is that they require a complete data tensor, which has to be built using labels for each mode of variation. That is, in the aforementioned example of faces with varying expression, identity and pose, one needs facial images for every possible expression and pose for each and every person in order to build the required complete tensor[1]. Clearly, this requirement limits the applicability of multilinear decompositions to data captured in controlled conditions (e.g., PIE [12], Multi-PIE [13] and BU-3DFE [14]), where all the necessary data variations along with their labels are available.

In this paper, we investigate the problem of disentangling the modes of variation in unlabelled and possibly incomplete data. In particular, we focus on sets of data that are incomplete in the sense that access to samples exhibiting every possible type of variation is not guaranteed. To this end, we propose the first *unsupervised multilinear decomposition* which uncovers the potential multilinear structure of incomplete sets of data and the corresponding low-dimensional latent variables (coefficients) explaining different types of variation. The proposed model is schematically summarised in Figure 1. In the depicted example, each image $\mathbf{x}_i$ is generated as tensor to vector product [6] of a tensor $\boldsymbol{\mathcal{B}}$ capturing the multilinear structure of the data and some coefficients corresponding to a meaningful variation. Here, $\mathbf{l}_i$ represents lighting coefficients, $\mathbf{e}_i$ expression coefficients and $\mathbf{c}_i$ identity coefficients. The number of differed types of variation is assumed to be known and specifies the order of the multilinear basis $\boldsymbol{\mathcal{B}}$.

The contributions here significantly extend the preliminary version of the paper [15] and are organized as follows:

- *M. Wang, Y. Panagakis and S. Zafeiriou are with the Dept. of Computing, Imperial College London, UK. Y. Panagakis is also with the Dept. of Computer Science, Middlesex University London, UK*
  Corresponding author: M. Wang, email: m.wang15@imperial.ac.uk
- *P. Snape was with the with the Dept. of Computing, Imperial College London, UK. He is now with Apple.*

Manuscript received March 1, 2017.

1. Methods for completing the tensor have been proposed but they are only approximate [9], [10], [11].

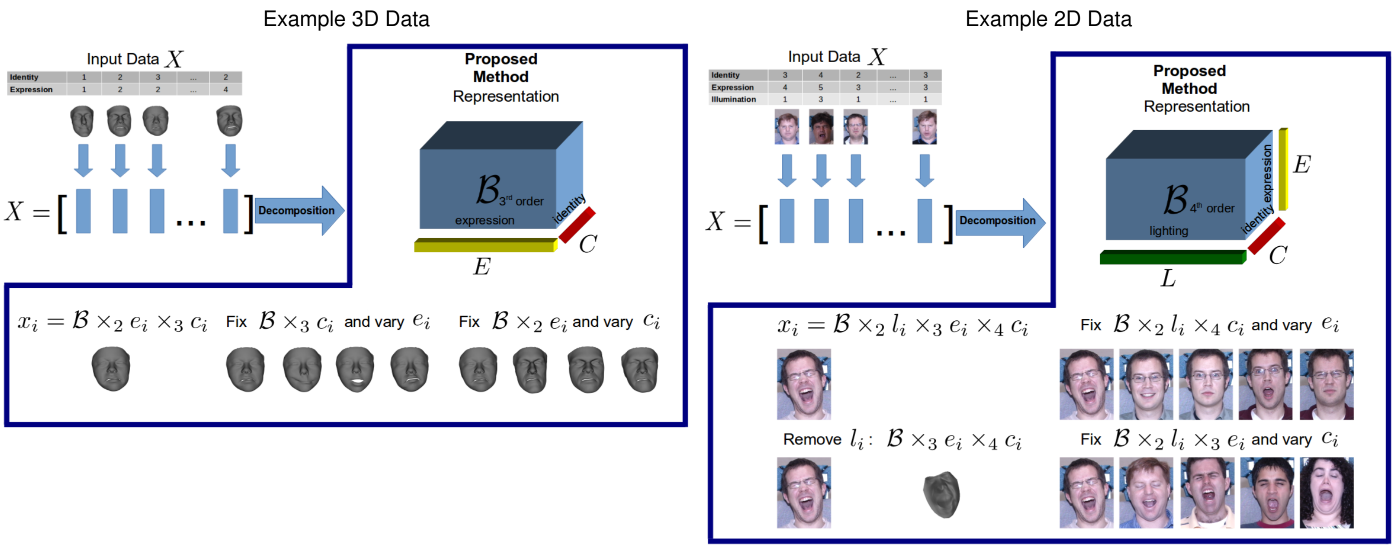


Fig. 1. Visualisation of the unsupervised multilinear decomposition and its applications. A sample vector $\mathbf{x}_i$ is assumed to be generated by a common multilinear structure $\mathcal{B}$ and sample specific weights e.g. $\mathbf{l}_i$, $\mathbf{e}_i$ and $\mathbf{c}_i$. We assume the weights correspond to variations in the data ($\mathbf{l}_i$ to lighting, $\mathbf{e}_i$ to expression and $\mathbf{c}_i$ to identity). By varying $\mathbf{e}_i$ only, we expect to see changes in expression but no change in identity or lighting. Similarly, if we vary $\mathbf{c}_i$ only we expect the expression and lighting to remain the same but the identities to change. Additionally if we remove the lighting $\mathbf{l}_i$, we expect the remaining information to correspond to the 3D shape of the object.

- A novel multilinear decomposition of matrices that recovers the multilinear structure and hence disentangles an arbitrary number of different modes of variation from possibly incomplete set of (visual) data is proposed in Section 4.1.
- To compute the proposed multilinear matrix decomposition, an efficient alternating least squares type of algorithm is developed in Section 4.1.
- The proposed method is extended to handle data contaminated by sparse noise of large magnitude and outliers in Section 4.2. To this end, a suitable ℓ_1 -norm regularized problem is solved allowing the estimation of different modes of variation in the presence of noise.
- A second variant of the proposed decomposition allowing the estimation of low-rank latent coefficients is introduced in Section 4.3. Latent coefficients with low-rank structure naturally appear in applications such as video analysis where consecutive video frames are highly correlated
- In practice, partial information regarding labels or the geometry of a subset of modes of variation is available. To exploit such information a graph-regularized extension of the proposed decomposition is proposed in Section 4.4.
- To demonstrate the generality of the proposed models, in Section 5 extensive experiments on computer vision tasks are conducted including facial expression transfer, Shape from Shading (SfS), and estimation of surface normals directly from "in-the-wild" images. In the latter task, we demonstrate that by feeding the estimated normals from the proposed decomposition into a deep neural network, facial reconstruction can be achieved using a single non-aligned image captured in the wild. Furthermore, it is worth mentioning that the methods for SfS in [16], [17] are only special cases of the proposed multilinear decomposition.

2 Notations and Multilinear Algebra Basics

Throughout the paper, matrices (vectors) are denoted by uppercase (lowercase) boldface letters e.g., $\mathbf{X}$, ($\mathbf{x}$). $\mathbf{I}$ denotes the identity matrix of compatible dimensions. The ith column of $\mathbf{X}$ is denoted as $\mathbf{x}_i$. Tensors are considered as the multidimensional equivalent of matrices (second-order tensors) and vectors (first-order tensors) and denoted by calligraphic letters, e.g., $\mathcal{X}$. The *order* of a tensor is the number of indices needed to address its elements. Consequently, each element of an Mth-order tensor $\mathcal{X}$ is addressed by M indices, i.e., $(\mathcal{X})_{i_1,i_2,\ldots,i_M} \doteq x_{i_1,i_2,\ldots,i_M}$.

The sets of real and integers numbers is denoted by $\mathbb{R}$ and $\mathbb{Z}$, respectively. A set of M real matrices (vectors) of varying dimensions is denoted by $\{\mathbf{X}^{(m)} \in \mathbb{R}^{I_n \times N}\}_{m=1}^{N}$ ($\{\mathbf{x}^{(m)} \in \mathbb{R}^{I_m}\}_{m=1}^{M}$). An Mth-order real-valued tensor $\mathcal{X}$ is defined over the tensor space $\mathbb{R}^{I_1 \times I_2 \times \cdots \times I_M}$, where $I_m \in \mathbb{Z}$ for $m = 1, 2, \ldots, M$.

An Mth-order tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_M}$ has *rank*-1, when it is decomposed as the outer product of M vectors $\{\mathbf{x}^{(m)} \in \mathbb{R}^{I_m}\}_{m=1}^{M}$. That is, $\mathcal{X} = \mathbf{x}^{(1)} \circ \mathbf{x}^{(2)} \circ \cdots \circ \mathbf{x}^{(M)} \doteq \bigcirc_{m=1}^{M} \mathbf{x}^{(m)}$, where $\circ$ denotes for the vector outer product.

The *mode-m matricisation* of a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_M}$ maps $\mathcal{X}$ to a matrix $\mathbf{X}_{(m)} \in \mathbb{R}^{I_m \times \bar{I}_m}$ with $\bar{I}_m = \prod_{k=1, k \neq m}^{M} I_k$ such that the tensor element $x_{i_1,i_2,\ldots,i_M}$ is mapped to the matrix element $x_{i_m,j}$ where $j = 1 + \sum_{k=1, k \neq m}^{M} (i_k - 1) J_k$ with $J_k = \prod_{n=1, n \neq m}^{k-1} I_n$.

The *mode-m vector product* of a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_M}$ with a vector $\mathbf{x} \in \mathbb{R}^{I_m}$, denoted by $\mathcal{X} \times_n \mathbf{x} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_{n-1} \times I_{n+1} \times \cdots \times I_N}$. The result is of order $M - 1$ and is defined element-wise as

$$(\mathcal{X} \times_m \mathbf{x})_{i_1,\ldots,i_{m-1},i_{m+1},\ldots,i_M} = \sum_{i_m=1}^{I_m} x_{i_1,i_2,\ldots,i_M} x_{i_m}. \tag{1}$$

In order to simplify the notation, we denote $\mathcal{X} \times_1 \mathbf{x}^{(1)} \times_2 \mathbf{x}^{(2)} \times_3 \cdots \times_M \mathbf{x}^{(M)} = \mathcal{X} \prod_{m=1}^{m} \times_m \mathbf{x}^{(m)}$.

The *Khatri-Rao* (column-wise Kronecker product) product of matrices $\mathbf{A} \in \mathbb{R}^{I \times N}$ and $\mathbf{B} \in \mathbb{R}^{J \times N}$ is denoted by $\mathbf{A} \odot \mathbf{B}$ and yields a matrix of dimensions $(IJ) \times N$. Furthermore, the Khatri-Rao of a set of matrices $\{\mathbf{X}^{(m)} \in \mathbb{R}^{I_m \times N}\}_{m=1}^{N}$ is denoted by $\mathbf{X}^{(1)} \odot \mathbf{X}^{(2)} \odot \cdots \odot \mathbf{X}^{(M)} \doteq \bigodot_{m=1}^{M} \mathbf{X}^{(m)}$. More details on tensors and multilinear operators can be found in [18] for example.

Finally, $\|\cdot\|_F$ denotes the Frobenius norm, $\|\cdot\|_*$ the nuclear norm and $\|\cdot\|_1$ the l_1-norm.

3 Related Works

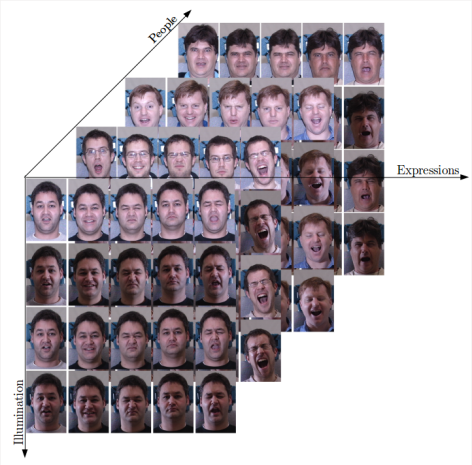


Fig. 2. Visualisation of the Multi-PIE [13] dataset. Collecting data where every person is present in all the lighting and expression variations is an expensive process that does not scale well.

For the past fifteen years, the computer vision community has made considerable efforts to collect databases in controlled conditions that can capture the variations of visual objects such as human faces. Arguably, the most comprehensive efforts were made in order to collect the so-called PIE [12] and Multi-PIE [13] databases. These databases contain a number of people (i.e., multiple identities) captured under different poses and illuminations, displaying a variety of facial expressions. Thus, this data sets contain many different modes of variation. These datasets motivated the use of multilinear decompositions, such as HOSVD [4], in order to disentangle the different modes of variations. The TensorFaces [8] is probably the most popular method in this category.

Concretely, let $\mathcal{X}$ be a complete data tensor (see Fig 2), TensorFaces [8] disentangle the modes of variation by seeking a decomposition of the form;

$$\mathcal{X} = \mathcal{B} \times_1 \boldsymbol{A}_1 \times_2 \boldsymbol{A}_2 \times_3 \boldsymbol{A}_3 \cdots \times_N \boldsymbol{A}_N, \tag{2}$$

where $\mathcal{B}$ is the core tensor of the same size as $\mathcal{X}$ representing the interactions between the factors $\boldsymbol{A}_n$, for $n = 1, \ldots, N$. Equation (2) directly corresponds to the HOSVD [4] formulation.

For example, on the Weizmann face dataset of 28 subjects in 5 viewpoints, 4 illuminations, 3 expressions and 7943 pixels per image, $\mathcal{X}$ is a $28 \times 5 \times 4 \times 3 \times 7943$ tensor. The aim is then to decompose $\mathcal{X}$ as

$$\mathcal{X} = \mathcal{B} \times_1 \boldsymbol{A}_{people} \times_2 \boldsymbol{A}_{views} \times_3 \boldsymbol{A}_{illumns} \times_4 \boldsymbol{A}_{expres} \times_5 \boldsymbol{A}_{pixels}, \tag{3}$$

where $\mathcal{B}$ is the $28 \times 5 \times 4 \times 3 \times 7943$ core tensor. [8] proposes the following N-mode SVD algorithm to recover this representation:

1) For $n \in \{people, views, illumns, expres, pixels\}$, flatten $\mathcal{X}$ into the matrix $\boldsymbol{X}_{(n)}$ and compute the SVD: $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T = SVD(\boldsymbol{X}_{(n)})$. Then set $\boldsymbol{A}_n = \boldsymbol{U}$.
2) Solve for $\mathcal{B}$ as: $\mathcal{B} = \mathcal{X} \times_1 \boldsymbol{A}_{people}^T \times_2 \boldsymbol{A}_{views}^T \times_3 \boldsymbol{A}_{illumns}^T \times_4 \boldsymbol{A}_{expres}^T \times_5 \boldsymbol{A}_{pixels}^T$

Even though TensorFaces and related methods e.g., [19] succeed in recovering the modes of variation, their applicability is rather limited since they not only require the data to be labelled but also the data tensor must contain all samples in all different variations. This is the primary reason that such methods are still mainly applied to tightly controlled databases such as PIE and Multi-PIE, visualised in Figure 2, and not to possibly not complete data captured "in-the-wild" data

A seemingly unrelated area of research that relies heavily on data decomposition is that of facial SfS [20] and Uncalibrated Photometric Stereo in General Lighting (UPS) [21]. The recovery of 3D shape from images represents an ill-posed and challenging problem. In its most difficult form, this involves recovering a representation of shape for an object from a single image, under arbitrary illumination. However, for any given image, there are an infinite number of shape, illumination and reflectance inputs that can reproduce the image [22]. Therefore, shape recovery is commonly performed by relaxing the problem by introducing prior information or by adding constraints, such as in SfS [20]. In particular, Class-specific UPS seeks to recover the shape of the object by exploiting the similarity within the object class. In the case of faces, there are millions of available images that can be utilised to build in-the-wild models. However, recovering shape from these images is incredibly challenging, as they have been captured in completely unconstrained conditions. No knowledge of the lighting conditions, the facial location or the camera geometric properties are provided with the images.

Recent class-specific UPS techniques [16], [17] proposed to recover a class-specific spherical harmonic (SH) basis that exploits the low-rank structure of faces [23], [24]. Spherical harmonics are ideal for this purpose as they can be approximated by a low-dimensional linear subspace [23], [25]. By using the first order SH, 87.5% of the low-frequency component of the lighting is approximated. The first order SH can then be used to recover 3D shape as their discrete approximation directly incorporates the normals of the object. These normals can be integrated to provide a dense 3D surface [26].The recovered SH basis can be robustly learnt from automatically aligned, "in-the-wild" images. [16], [17] attempt to build a subspace that explicitly separates shape and appearance by performing a rank constrained Khatri-Rao (KR) factorization [27]. The first paper where the decomposition has been proposed and applied in 3D facial shape reconstruction is [16]. The method in [16] was inspired by the decomposition techniques employed in the related area of Structure-from-Motion [28]. Recently, [17] made the link between the KR factorisation and the UPS and utilised the method by [29] to solve the KR factorisation. [17] proposed a robust decomposition in place of the

optical flow [30] based registration used by [16] to remove outliers from the images. In this paper, we show that the decompositions proposed in [16], [17] are very special cases of the proposed unsupervised tensor decomposition.

To alleviate the aforementioned limitations and disentangle an arbitrary number of modes of variation without having labels and complete data a novel multilinear matrix decomposition as well as several extensions are presented in detail next.

4 Proposed Methods

4.1 Basic Model

Let $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N] \in \mathbb{R}^{d \times N}$ be a matrix of observations, where each of the N columns represent a vectorised image of d pixels. In order to discover $M-1$ different modes of variation we propose the following decomposition:

$$\boldsymbol{x}_i = \boldsymbol{\mathcal{B}} \times_2 \boldsymbol{a}_i^{(2)} \times_3 \boldsymbol{a}_i^{(3)} \cdots \times_m \boldsymbol{a}_i^{M} = \boldsymbol{\mathcal{B}} \prod_{m=2}^{M} \times_m \mathbf{a}_i^{(m)}, \tag{4}$$

where $\boldsymbol{\mathcal{B}} \in \mathbb{R}^{d \times K_2 \times \cdots \times K_M}$ representing the common multilinear basis of $\boldsymbol{X}$ and the set of vectors $\{\boldsymbol{a}_i^{(m)} \in \mathbb{R}^{K_m}\}_{m=2}^{M}$ represents the variation coefficients in each mode specific to the vectorised image $\boldsymbol{x}_i$.

Therefore, for the observation matrix $\boldsymbol{X}$, and by exploiting the properties of multilinear operators e.g., [18], the above decomposition is written in matrix form as

$$\boldsymbol{X} = \boldsymbol{B}_{(1)}(\boldsymbol{A}^{(2)} \odot \boldsymbol{A}^{(3)} \cdots \odot \boldsymbol{A}^{(M)}) = \boldsymbol{B}_{(1)}\big(\bigodot_{m=2}^{M} \mathbf{A}^{(m)}\big), \tag{5}$$

where $\boldsymbol{B}_{(1)}\mathbb{R}^{d \times K_2 \cdot K_3 \cdot K_M}$ is the mode-1 matricisation of $\boldsymbol{\mathcal{B}}$ and $\{\boldsymbol{A}^{(m)}\}_{m=2}^{M} \in \mathbb{R}^{K_m \times N}$ gathers the variation coefficients for all images across $M-1$ modes of variation. Clearly, this formulation is different from the Tucker decomposition [3] and the HOSVD [4].

To find the unknown multilinear basis $\boldsymbol{\mathcal{B}}$ and the variation coefficients $\{\boldsymbol{A}^{(m)}\}_{m=2}^{M}$, we propose to solve:

$$\begin{aligned} \underset{\boldsymbol{B}_{(1)}, \{\boldsymbol{A}^{(m)}\}_{m=2}^{M}}{\arg\min} \; & \|\boldsymbol{X} - \boldsymbol{B}_{(1)}\big(\bigodot_{m=2}^{M} \mathbf{A}^{(m)}\big)\|_F^2 \\ \text{s.t.} \; & \boldsymbol{B}_{(1)}^T \boldsymbol{B}_{(1)} = \mathbf{I}. \end{aligned} \tag{6}$$

Optimisation problem (6) is non-convex. Therefore, we propose to solve (6) by employing an Alternating Least Squares (ALS) scheme, where each variable is updated in an alternating fashion. Let t denotes the iteration index, given $\boldsymbol{B}_{(1)}[0]$ and $\{\boldsymbol{A}^{(m)}[0]\}_{m=2}^{M}$, the iteration of the ALS solver reads as follows:

$$\begin{aligned} \boldsymbol{B}_{(1)}[t+1] = \underset{\boldsymbol{B}_{(1)}}{\arg\min} \; & \|\boldsymbol{X} - \boldsymbol{B}_{(1)}\big(\bigodot_{m=2}^{M} \mathbf{A}^{(m)}[t]\big)\|_F^2 \\ \text{s.t.} \; & \boldsymbol{B}_{(1)}^T \boldsymbol{B}_{(1)} = \mathbf{I}. \end{aligned} \tag{7}$$

$$\{\boldsymbol{A}^{(m)}[t+1]\}_{m=2}^{M} = \underset{\{\boldsymbol{A}^{(m)}\}_{m=2}^{M}}{\arg\min} \|\boldsymbol{X} - \boldsymbol{B}_{(1)}[t+1]\big(\bigodot_{m=2}^{M} \mathbf{A}^{(m)}\big)\|_F^2 \tag{8}$$

Solving (7): Problem (7) is an orthogonal Procrustes problem, whose solution is given by [31]: $\boldsymbol{B}_{(1)}[t+1] = \boldsymbol{U}\boldsymbol{V}^T$, where $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T = \boldsymbol{X}\big(\bigodot_{m=2}^{M} \mathbf{A}^{(m)}[t]\big)^T$ is the SVD.

Solving (8): Due to the unitary invariance of the Frobenius norm (7) is equivalent to

$$\underset{\{\boldsymbol{A}^{(m)}\}_{m=2}^{M}}{\arg\min} \|\boldsymbol{B}_{(1)}[t+1]^T \boldsymbol{X} - \bigodot_{m=2}^{M} \mathbf{A}^{(m)}\|_F^2, \tag{9}$$

which is a Khatri-Rao factorisation problem [29]. Let $\boldsymbol{Q} = \boldsymbol{B}_{(1)}[t+1]^T \boldsymbol{X} \in \mathbb{R}^{K_2 \cdot K_3 \cdots K_M \times N}$, then each column of Q is written as:

$$\boldsymbol{q}_i = \bigodot_{m=2}^{M} \mathbf{a}_i^{(m)} \tag{10}$$

Let us partition $\boldsymbol{q}_i$ into a set $S = K_{M-1} \cdot K_{M-2} \cdots K_2$ vectors $\{\boldsymbol{q}_i^{B_b} \in \mathbb{R}^{K_M}\}_{b=1}^{K_{M-1} \cdot K_{M-2} \cdots K_2}$ such that $\boldsymbol{q}_i = [\boldsymbol{q}_i^{B_1 T} \boldsymbol{q}_i^{B_2 T} \cdots \boldsymbol{q}_i^{B_S T}]^T$. This partitioning enables us to rearranging the elements of $\boldsymbol{q}_i$ into a tensor $\boldsymbol{Q}_i \in \mathbb{R}^{K_M \times K_{M-1} \times \cdots \times K_2}$ such that $\boldsymbol{Q_i}_{(1)} = [\boldsymbol{q}_i^{B_1}, \boldsymbol{q}_i^{B_2}, \cdots, \boldsymbol{q}_i^{B_{K_{M-1} \cdot K_{M-2} \cdots K_2}}] \in \mathbb{R}^{K_M \times (K_{M-1} \cdot K_{M-2} \cdots K_2)}$. Therefore, based on (10), $\boldsymbol{Q}_i$ is written as

$$\boldsymbol{Q}_i = \mathbf{a}_i^{(M)} \circ \mathbf{a}_i^{(M-1)} \circ \cdots \circ \mathbf{a}_i^{(2)} \tag{11}$$

Equation (11) indicates that we can recover the set of vectors $\{\mathbf{a}_i^{(m)}\}_{m=2}^{M}$ and therefore the set of matrices $\{\mathbf{A}^{(m)}\}_{m=2}^{M}$, by seeking a best (in the least squares sense) rank-1 approximation of $\boldsymbol{Q}_i$, for $i = 1, 2, \ldots, N$. An efficient way to find the best rank-1 approximation of $\boldsymbol{Q}_i$ is to exploit the truncated HOSVD [4]. That is,

$$\boldsymbol{Q}_i = s \prod_{n=1}^{M-1} \times_n \mathbf{u}_i^{(n)}, \tag{12}$$

where $\{\mathbf{u}_i^{(n)} \in \mathbb{R}^{K_{M-n+1}}\}_{n=1}^{M-1}$ is the the set of the first higher order singular vector along $M-1$ modes of tensor $\boldsymbol{Q}_i$ and $s = (\boldsymbol{\mathcal{S}})_{1,1,\ldots,1}$ is the first high-order singular value stored as a first element in the core tensor $\boldsymbol{\mathcal{S}}$. Consequently, the columns of the variation coefficient matrices $\{\mathbf{A}^{(m)}\}_{m=2}^{M}$ can be estimated by

$$\boldsymbol{a}_i^{(m)} = s^{\frac{1}{M-1}} \mathbf{u}_i^{(M-m+1)}, \tag{13}$$

for $m = 2, 3, \ldots, M$. Interestingly, the estimation of the variation coefficients according to (13) resolves the inherent scaling ambiguity in (9) by assigning the same Euclidean-norm to each column of $\mathbf{A}^{(m)}$. The procedure of solving (6) is summarised in Algorithm 1.

Remarks: In the special case of 2 modes and where $k_2 = 4$, (5) becomes:

$$\boldsymbol{X} = \boldsymbol{B}_{(1)}(\boldsymbol{L} \odot \boldsymbol{C}), \tag{14}$$

where $\boldsymbol{L} = \boldsymbol{A}_2 \in \mathbb{R}^{4 \times n}$, $\boldsymbol{C} = \boldsymbol{A}_3 \in \mathbb{R}^{k \times n}$.
Let $\boldsymbol{P} = \boldsymbol{L} \odot \boldsymbol{C}$ then,

$$\boldsymbol{X} = \boldsymbol{B}_{(1)}\boldsymbol{P}. \tag{15}$$

Equation (15) corresponds to the formulation used by [16]. $\boldsymbol{P} = \boldsymbol{L} \odot \boldsymbol{C}$ has been implied by [16] but not explicitly formulated as such. Hence this shows that [16] represents a special case of our general decomposition.

4.2 Robust Decomposition

Equation (14) corresponds to the formulation used by [17] with the only difference being the separation of $\boldsymbol{X}$ into a low-rank part $\boldsymbol{BP}$ and sparse, non-Gaussian error $\boldsymbol{E}$. [17] used these to add robustness to the decomposition due to occlusion that is present in images captured in unconstrained conditions (also referred to as "in-the-wild"), as well as to account for the high frequency errors introduced by the coarse geometric alignment of the images. Generalising

Algorithm 1 Multilinear Data Decomposition Algorithm

Input: Data Matrix $\boldsymbol{X} \in \mathbb{R}^{d \times N}$, dimensions $K_2, K_3, \dots K_M$
Output: $\mathcal{B}$, $\boldsymbol{A}^{(2)}$, $\boldsymbol{A}^{(3)}, \dots \boldsymbol{A}^{(M)}$

Initialisation: $t \leftarrow 0$,
$[\boldsymbol{U}, \boldsymbol{\Sigma}, \boldsymbol{V}] \leftarrow$SVD$(\boldsymbol{X})$,
$\boldsymbol{B_{(1)}}[0] = \boldsymbol{U}\sqrt{\boldsymbol{\Sigma}}$, $\boldsymbol{Q}[0] = \sqrt{\boldsymbol{\Sigma}}\boldsymbol{V}^T$
while not converged **do**
  for all image $i = 1 \dots N$ **do**
    construct $\boldsymbol{Q}_i \in \mathbb{R}^{K_M \times K_{M-1} \cdots \times K_2}$ from $\boldsymbol{q}_i[t]$
    $[\boldsymbol{S}_i, \boldsymbol{U}_i] \leftarrow$HOSVD$(\boldsymbol{Q}_i)$
    for each mode m = 2 ... M **do**
      $\boldsymbol{a}_i^{(m)}[t+1] = (\boldsymbol{S}_i)_1^{\overline{M-1}} \boldsymbol{U}_i^{(M-m+1)}$
    end for
  end for
  $[\boldsymbol{U}, \boldsymbol{\Sigma}, \boldsymbol{V}] \leftarrow$SVD$(\boldsymbol{X}(\bigodot_{m=2}^{M} \mathbf{A}^{(m)}[t])^T)$
  $\boldsymbol{B_{(1)}}[t+1] = \boldsymbol{U}\boldsymbol{V}^T$
  $\boldsymbol{Q}[t+1] = \boldsymbol{B}_{(1)}[t+1]^T \boldsymbol{X}$
  Check convergence condition:
  $\frac{\|\boldsymbol{X} - \boldsymbol{B_{(1)}}[t+1]\boldsymbol{Q}[t+1]\|_F^2}{\|\boldsymbol{X}\|_F^2} < \epsilon$
  $t \leftarrow t+1$
end while
Tensorise $\boldsymbol{B_{(1)}}$ into $\mathcal{B} \in \mathbb{R}^{d \times K_2 \times \cdots \times K_M}$

this to the case of arbitrary number of modes, a robust decomposition can be found by solving:

$$\begin{aligned} \underset{\boldsymbol{B_{(1)}}, \{\boldsymbol{A}^{(m)}\}_{m=2}^{M}}{\arg\min} \quad & \|\boldsymbol{P}\|_* + \lambda \|\boldsymbol{E}\|_1 \\ \text{s.t.} \quad & \boldsymbol{X} = \boldsymbol{B_{(1)}}\boldsymbol{P} + \boldsymbol{E}, \\ & \boldsymbol{P} = (\bigodot_{m=2}^{M} \mathbf{A}^{(m)}), \\ & \boldsymbol{B}_{(1)}^T \boldsymbol{B_{(1)}} = \mathbf{I}. \end{aligned} \tag{16}$$

The nuclear norm and the l_1-norm promote low-rank and sparsity respectively. To solve (16), the Alternating Directions Method (ADM) [32] is employed. To this end, the following augmented Lagrangian function should be minimised:

$$\begin{aligned} &\mathcal{L}(\boldsymbol{P}, \boldsymbol{E}, \boldsymbol{B_{(1)}}, \{\boldsymbol{A}^{(m)}\}_{m=2}^{M}, \boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2) = \\ &\|\boldsymbol{P}\|_* + \lambda\|\boldsymbol{E}\|_1 + \frac{\mu}{2}\|\boldsymbol{X} - \boldsymbol{B_{(1)}}\boldsymbol{P} - \boldsymbol{E} + \frac{\boldsymbol{\Lambda}_1}{\mu}\|_F^2 + \\ &\frac{\mu}{2}\|\boldsymbol{P} - (\bigodot_{m=2}^{M} \mathbf{A}^{(m)}) + \frac{\boldsymbol{\Lambda}_2}{\mu}\|_F^2, \end{aligned} \tag{17}$$

with respect to $\boldsymbol{B}_{(1)}^T \boldsymbol{B_{(1)}} = \mathbf{I}$. $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\Lambda}_2$ denote the Lagrangian multipliers.

An ADM-Solver for (17) can be found in Algorithm 2, while its derivation can be found in the supplementary material.

4.3 Rank-constrained Decomposition

We propose another variation to the problem formulated in (6) in order to incorporate additional low-rank constraints so that the methodology is suitable for image analysis. A sequence of images of a face from a single viewpoint, under varying illumination, can be nearby completely explained by 5 or 6 principal components [33]. This justifies the addition of an additional low-rank constraint in the case of specific data such as videos of a single person under illumination change.

Algorithm 2 Robust Multilinear Decomposition Algorithm

Input: Data Matrix $\boldsymbol{X} \in \mathbb{R}^{d \times N}$, dimensions $K_2, K_3, \dots K_M$ and parameter λ **Output:** $\mathcal{B}$, $\boldsymbol{E}$, $\boldsymbol{P}$, $\boldsymbol{A}^{(2)}$, $\boldsymbol{A}^{(3)}, \dots \boldsymbol{A}^{(M)}$

Initialisation:
$t \leftarrow 0$,
$\boldsymbol{P}[0] = 0$, $\boldsymbol{E}[0] = 0$, $\boldsymbol{B_{(1)}}[0] = 0$, $\{\boldsymbol{A}^{(m)} = 0\}_{m=2}^{M}$,
$\boldsymbol{\Lambda_1}[0] = 0$, $\boldsymbol{\Lambda_2}[0] = 0$, $\mu = 10^{-6}$, $\rho = 1.1$, $\epsilon = 10^{-8}$
while not converged **do**
  Update $\boldsymbol{P}[t+1]$ by

$$\boldsymbol{P}[t+1] = \mathcal{D}_{\mu^{-1}}\big[\boldsymbol{B_{(1)}}[t]^T\boldsymbol{X} - \boldsymbol{P}[t] - \boldsymbol{B_{(1)}}[t]^T\boldsymbol{E} + \frac{\boldsymbol{B_{(1)}}[t]^T\boldsymbol{\Lambda_1}[t]}{\mu} + (\bigodot_{m=2}^{M} \mathbf{A}^{(m)}[t]) - \frac{\boldsymbol{B_{(1)}}[t]^T\boldsymbol{\Lambda_2}[t]}{\mu}\big]$$

  Update $\boldsymbol{E}[t+1]$ by

$$\boldsymbol{E}[t+1] = \mathcal{S}_{\lambda\mu^{-1}}\big[\boldsymbol{X} - \boldsymbol{B_{(1)}}[t]\boldsymbol{P}[t+1] + \frac{\boldsymbol{\Lambda_1}[t]}{\mu}\big]$$

  Update $\boldsymbol{B_{(1)}}[t+1]$ by

$$(\boldsymbol{X} - \boldsymbol{E}[t+1] + \frac{\boldsymbol{\Lambda_1}[t]}{\mu})\boldsymbol{P}[t+1]^T = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T,$$
$$\boldsymbol{B_{(1)}}[t+1] = \boldsymbol{U}\boldsymbol{V}^T$$

  Set $\boldsymbol{Q} = \boldsymbol{P}[t+1] + \frac{\boldsymbol{\Lambda}_2[t]}{\mu}$
  for all image $i = 1 \dots N$ **do**
    construct $\boldsymbol{Q}_i \in \mathbb{R}^{K_M \times K_{M-1} \cdots \times K_2}$ from $\boldsymbol{q}_i[t]$
    $[\boldsymbol{S}_i, \boldsymbol{U}_i] \leftarrow$HOSVD$(\boldsymbol{Q}_i)$
    for each mode m = 2 ... M **do**
      $\boldsymbol{a}_i^{(m)}[t+1] = (\boldsymbol{S}_i)_1^{\overline{M-1}} \boldsymbol{U}_i^{(M-m+1)}$
    end for
  end for
  Update Lagrange multipliers by

$$\boldsymbol{\Lambda_1}[t+1] = \boldsymbol{\Lambda_1}[t] + \mu(\boldsymbol{X} - \boldsymbol{B_{(1)}}[t+1]\boldsymbol{P}[t+1] - \boldsymbol{E}[t+1])$$
$$\boldsymbol{\Lambda_2}[t+1] = \boldsymbol{\Lambda_2}[t] + \mu(\boldsymbol{P}[t+1] - (\bigodot_{m=2}^{M} \mathbf{A}^{(m)}[t+1]))$$

  Update μ by $\mu = min(\rho\mu, 10^{-6})$
  Check convergence condition:

$$\frac{\|\boldsymbol{X} - \boldsymbol{B_{(1)}}[t+1]\boldsymbol{P}[t+1] - \boldsymbol{E}[t+1]\|_F^2}{\|\boldsymbol{X}\|_F^2} < \epsilon$$
$$\frac{\|\boldsymbol{P}[t+1] - (\bigodot_{m=2}^{M} \mathbf{A}^{(m)}[t+1])\|_F^2}{\|\boldsymbol{X}\|_F^2} < \epsilon$$

  $t \leftarrow t+1$
end while
Tensorise $\boldsymbol{B_{(1)}}$ into $\mathcal{B} \in \mathbb{R}^{d \times K_2 \times \cdots \times K_M}$

We propose to solve the following problem:

$$\begin{aligned} \underset{\boldsymbol{B}_{(1)},\{\boldsymbol{A}^{(m)}\}_{m=2}^{M}}{\arg\min} \quad & \{\|\boldsymbol{A}^{(m)}\|_*\}_{m=2}^{M} + \lambda\|\boldsymbol{E}\|_1 \\ \text{s.t.} \quad & \boldsymbol{X} = \boldsymbol{B}_{(1)}\big(\bigodot_{m=2}^{M} \mathbf{L}^{(m)}\big) + \boldsymbol{E}, \\ & \boldsymbol{L}^{(m)} = \boldsymbol{A}^{(m)}, \\ & \boldsymbol{B}_{(1)}^T\boldsymbol{B}_{(1)} = \mathbf{I}. \end{aligned} \tag{18}$$

Similarly to (16), Problem (18) is solved by employing the ADM. That is, the following augmented Lagrangian function is minimised:

$$\begin{aligned} &\mathcal{L}(\{\boldsymbol{A}^{(m)}\}_{m=2}^{M}, \boldsymbol{E}, \boldsymbol{B}_{(1)}, \{\boldsymbol{\Lambda}^{(m)}\}_{m=1}^{M}) = \\ &\{\|\boldsymbol{A}^{(m)}\|_*\}_{m=2}^{M} + \lambda\|\boldsymbol{E}\|_1 + \\ &\frac{\mu}{2}\|\boldsymbol{X} - \boldsymbol{B}_{(1)}\big(\bigodot_{m=2}^{M} \mathbf{L}^{(m)}\big) - \boldsymbol{E} + \frac{\boldsymbol{\Lambda}_1}{\mu}\|_F^2 + \\ &\sum_{m=2}^{M} \frac{\mu}{2}\|\boldsymbol{L}^{(m)} - \boldsymbol{A}^{(m)} + \frac{\boldsymbol{\Lambda}_m}{\mu}\|_F^2, \end{aligned} \tag{19}$$

with respect to $\boldsymbol{B}_{(1)}^T\boldsymbol{B}_{(1)} = \mathbf{I}$. $\boldsymbol{\Lambda}^{(m)}$ are the Lagrangian multipliers. Let t denote the iteration index.

An ADM-Solver for (19) can be found in Algorithm 3, while its derivation is similar to that of Algorithm 2 in Section 4.2 and it can be found in the supplementary material.

4.4 Graph-regularised Decomposition

In practical applications, there might be available side information about the geometric and topological properties of some modes of variation or even available labels. A typical example is a set of facial images with known identities captured under unknown illuminations conditions. To capture such geometric or label information, the graph embedding framework [34] can be employed by defining a suitable Laplacian graph capturing the available geometric or discriminant information. Therefore, a graph-regularized version of the proposed method is as follows:

$$\begin{aligned} \underset{\boldsymbol{B}_{(1)},\{\boldsymbol{A}^{(m)}\}_{m=2}^{M}}{\arg\min} \quad & \|\boldsymbol{X} - \boldsymbol{B}_{(1)}\big(\bigodot_{m=2}^{M} \mathbf{A}^{(m)}\big)\|_F^2 \\ & + \sum_{S\subseteq\{2,\ldots,M\}} \lambda_s tr(\boldsymbol{A}^{(s)}\boldsymbol{L}^{(s)}\boldsymbol{A}^{(s)T}) \\ \text{s.t.} \quad & \boldsymbol{B}_{(1)}^T\boldsymbol{B}_{(1)} = \mathbf{I}. \end{aligned} \tag{20}$$

$\boldsymbol{L}^{(s)} \in \mathbb{R}^{N\times N}$ corresponds to the Laplacian matrix containing either the labelled information or other constraints of a specific mode s. The λ_s are input weight parameters which balance the reconstruction error term with the graph constraints.

This decomposition can be applied in unsupervised, semi-supervised and supervised manner. Depending on the Laplacian matrix, different graph embeddings can be incorporated [34]. To apply this in a unsupervised manner, $\boldsymbol{L}^{(s)}$ can be specified to learn a manifold structure that conserves local structure such as used in Laplacianfaces [35]. In ISOMAP [36] the Laplacian is specified to preserve the geodesic distances of the data points.

Algorithm 3 Rank-constrained Multilinear Decomposition Algorithm

Input: Data Matrix $\boldsymbol{X} \in \mathbb{R}^{d\times N}$, dimensions K_2, K_3, $\ldots K_M$ and parameter λ **Output:** $\mathcal{B}$, $\boldsymbol{E}$, $\boldsymbol{A}^{(2)}$, $\boldsymbol{A}^{(3)}$, $\ldots \boldsymbol{A}^{(M)}$

1: Initialisation:
$t \leftarrow 0$,
$\boldsymbol{E}[0] = 0$, $\boldsymbol{B}_{(1)}[0] = 0$, $\{\boldsymbol{A}^{(m)} = 0\}_{m=2}^{M}$, $\{\boldsymbol{L}^{(m)} = 0\}_{m=2}^{M}$,
$\boldsymbol{\Lambda}_1[0] = 0$, $\{\boldsymbol{\Lambda}_m[0] = 0\}_{m=2}^{M}$, $\mu = 10^{-6}$, $\rho = 1.1$, $\epsilon = 10^{-8}$
2: **while** not converged **do**
3: Update $\{\boldsymbol{A}^{(m)}[t+1]\}_{m=2}^{M}$ by

$$\boldsymbol{A}^{(m)}[t+1] = \mathcal{D}_{\mu^{-1}}(\boldsymbol{L}^{(m)}[t] + \frac{\boldsymbol{\Lambda}_m[t]}{\mu})$$

4: Update $\boldsymbol{E}[t+1]$ by

$$\boldsymbol{E}[t+1] = \mathcal{S}_{\frac{\lambda}{\mu}}(\boldsymbol{X} - \boldsymbol{B}_{(1)}[t](\bigodot_{m=2}^{M} \mathbf{L}^{(m)}[t]) - \boldsymbol{E}[t] + \frac{\Lambda_1[t]}{\mu})$$

5: Set $\{\boldsymbol{O}_m = \boldsymbol{A}^{(m)}[t+1] - \frac{\boldsymbol{\Lambda}_m[t]}{\mu}\}_{m=2}^{M}$
6: **for all** image $i = 1 \ldots N$ **do**
7: **for** each mode s = 2 ... M **do**
8: $\boldsymbol{B'} = \mathcal{B}[t]\prod_{m=2, m\neq s}^{M} \times_m \mathbf{l}_i^{(m)}[t]$
9: $\mathbf{l}_i^{(s)}[t+1] = (2\boldsymbol{B'}^T\boldsymbol{B'} + \mu\boldsymbol{I})^{-1}(2\boldsymbol{B'}^T\boldsymbol{x}_i + \mu\boldsymbol{o}_i[t])$
10: **end for**
11: **end for**
12: Update $\boldsymbol{B}_{(1)}[t+1]$ by

$$(\boldsymbol{X} - \boldsymbol{E}[t+1] + \frac{\boldsymbol{\Lambda}_1[t]}{\mu})(\bigodot_{m=2}^{M} \mathbf{L}^{(m)}[t+1])^T = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T,$$
$$\boldsymbol{B}_{(1)}[t+1] = \boldsymbol{U}\boldsymbol{V}^T$$

13: Update Lagrange multipliers by

$$\boldsymbol{\Lambda}_1[t+1] = \boldsymbol{\Lambda}_1[t] + \mu(\boldsymbol{X} - \boldsymbol{B}_{(1)}[t+1](\bigodot_{m=2}^{M} \mathbf{L}^{(m)}[t+1]) - \boldsymbol{E}[t+1])$$
$$\{\boldsymbol{\Lambda}_m[t+1] = \boldsymbol{\Lambda}_m[t] + \mu(\boldsymbol{L}^{(m)}[t+1] - \mathbf{A}^{(m)}[t+1])\}_{m=2}^{M}$$

14: Update μ by $\mu = min(\rho\mu, 10^{-6})$
15: Check convergence condition:

$$\frac{\|\boldsymbol{X} - \boldsymbol{B}_{(1)}[t+1](\bigodot_{m=2}^{M} \mathbf{L}^{(m)}[t+1]) - \boldsymbol{E}[t+1]\|_F^2}{\|\boldsymbol{X}\|_F^2} < \epsilon$$
$$\{\frac{\|\boldsymbol{L}^{(m)}[t+1] - \mathbf{A}^{(m)}[t+1]\|_F^2}{\|\boldsymbol{X}\|_F^2} < \epsilon\}_{m=2}^{M}$$

16: $t \leftarrow t+1$
17: Tensorise $\boldsymbol{B}_{(1)}[t+1]$ into $\mathcal{B}[t+1] \in \mathbb{R}^{d\times K_2\times\cdots\times K_M}$
18: **end while**

In the case of semi-supervised learning, we specify $\boldsymbol{L}^{(s)}$ for specific modes s and include labelled information in $\boldsymbol{L}^{(s)}$ where available. For the samples where the labels are absent, we complete their entries in $\boldsymbol{L}^{(s)}$ by adding connections to their nearest neighbours. This type of semi-supervision has been previously used in [37]. For supervised learning, we can use the labels to form $\boldsymbol{L}^{(s)}$. The Laplacian can

also be specified as the graph embedding corresponding to LDA [38].

The algorithm and its derivation can be found in the supplementary material.

5 Experiments

In this section, we provide a number of experimental results in order to demonstrate the ability of the proposed method in recovering meaningful modes of variations. Unless otherwise stated, all the data used have been aligned to a reference shape to achieve pixel-wise correspondence.

- We first investigate synthetic 3D facial data that contains both facial expression and identity variations. As there is no texture variability, we provide a proof of concept of our methodology on 3D faces disentangling expression and identity.
- Then we consider 2D data captured in controlled conditions that simultaneously contains lighting, expression and identity variations. As a proof of concept, we show that the decomposition is able to disentangle the 3 different modes of variations. These experiments prove that the model is able to extract meaningful modes of variations from visual data.
- Thirdly we investigate "in-the-wild" datasets of faces and ears. Here, the l_1 optimisation is required to provide robustness as it is able to disregard noise and occlusions in the data. We robustly disentangle shape and illumination on the"in-the-wild" datasets. We also show that the robust decomposition is able to achieve better reconstruction results than the non-robust version. In a separate experiment, we consider "in-the-wild" videos of a single person. The use of the low-rank constraint allows us to disentangle shape and illumination and reconstruct the face despite synthetic or natural occlusions.
- We also show how our graph-regularised decomposition can be applied to disentangle expression and identity in a semi-supervised setting. We then test the resulting expression components for classification and find that they become more discriminative.
- Finally, we show that the low-rank subspace of shape we obtained from prior decompositions is extremely powerful. We create an unsupervised learning normal estimation pipeline in which we feed the estimated normals from our decomposition method on an "in-the-wild" dataset of faces as the input data to a deep neural network. The resulting deep network is then able to reconstruct faces from a single non-aligned "in-the-wild" image.

Overall, we demonstrate that our method requires neither complete well-organised data (e.g. all the objects under the same number of lighting conditions), nor labels to find the underlying multilinear structure. We also show that the extended decomposition methods can be applied to "in-the-wild" datasets of different objects to achieve superior performance.

5.1 Disentangling Expression and Identity

In this set of experiments we synthetically generate a dataset of 3D faces where the only variations are identity and expression. The dataset has been created using the Large Scale 3D Morphable Model [39] and put in correspondence with the blendshapes of the FaceWarehouse [40] so that we can allow for expressions. We used 200 components to describe identity and 10 components for expression. The dataset with 2000 3D facial meshes consists of 10 facial expressions and 200 identities. We wanted to examine whether our decomposition is able to find a space of identity variation that did not contain expressions. To this purpose we ensured that the facial expressions included in the data did not contain the neutral expression. A sample of the dataset is shown in Figure 4.

The decomposition becomes:

$$\boldsymbol{X} = \boldsymbol{B}_{(1)}(\boldsymbol{E} \odot \boldsymbol{C}), \tag{21}$$

where $\boldsymbol{B}_{(1)} \in \mathbb{R}^{d \times ek}$ is the orthogonal mode-1 matricisation of tensor $\boldsymbol{\mathcal{B}}$. $\boldsymbol{E} \in \mathbb{R}^{e \times n}$ is the matrix of expression coefficients. e should be set to the approximate number of differing expressions in the data. $\boldsymbol{C} \in \mathbb{R}^{k \times n}$ is assumed to be a matrix of identity coefficients. Evidently, this is a special case of our proposed decomposition in (5). The choice of k is subject to a trade-off between reconstruction detail of the data and the ability of the decomposition to separate expression and identity.

Given this setting and an appropriate choice for k, we performed a number of experiments to show that our decomposition is able to separate expression from identity. Setting $e = 10$ and $k = 50$, we apply the decomposition to discover that $\boldsymbol{\mathcal{B}} \in \mathbb{R}^{d \times e \times k}$ becomes a basis of expression and identity. We note that $\pm\boldsymbol{\mathcal{B}}_{:i:}$ are bases corresponding to expressions in the dataset. The first 10 components of the first 3 bases are plotted in Figure 3. We also discover that the first basis $\pm\boldsymbol{\mathcal{B}}_{:0:}$, visualised in Figure 3a, is a basis of neutral expressions. This is impressive as the neutral expression did not exist in the original dataset [2].

Thus we can use the neutral expression basis to create synthetic neutral faces of people using the following method. Let $\boldsymbol{B}_0$ denote the neutral expression basis $\boldsymbol{\mathcal{B}}_{:0:}$:

$$\boldsymbol{x}'_i = \boldsymbol{B}_0(\boldsymbol{B}_0^T\boldsymbol{B}_0)^{-1}\boldsymbol{B}_0^T\boldsymbol{x}_i, \tag{22}$$

where $\boldsymbol{x}'_i$ denotes the resulting neutral face of the person in $\boldsymbol{x}_i$. The results are visualised in Figure 5.

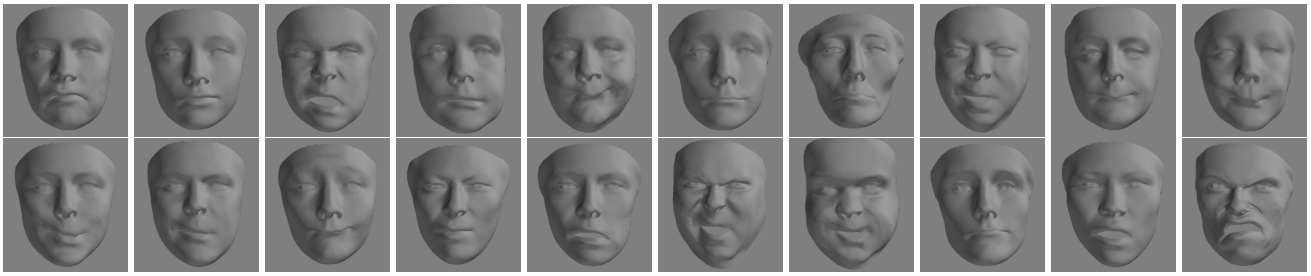
By decoupling $\boldsymbol{E}$, the matrix of expression coefficients and $\boldsymbol{C}$, the matrix representing identities, the decomposition allows us to transfer expressions across identities. Facial expression transfer results are in Figure 6.

5.2 Disentangling Illumination, Expression and Identity

In this experiment, we test our decomposition on data that simultaneously contains lighting, facial expression variations as well as multiple identities such as the Multi-PIE [13] dataset. We select 147 identities, 5 expressions

2. Nevertheless some, e.g. the 5th column of Figure 3 from the left do show some expression. This is mainly because we applied arbitrary scaling to the component (we could have normalized the scaling to the variance associated with this component). The experiments that can conclusively show that our method indeed decoupled identity and expression are shown in Figures 5 and 6 (transferring expression by changing only the components in $\boldsymbol{E}$).

(a) Basis of first expression $\pm\boldsymbol{\mathcal{B}}_{:0:}$



(b) Basis of second expression $\pm\boldsymbol{\mathcal{B}}_{:1:}$

(c) Basis of third expression $\pm\boldsymbol{\mathcal{B}}_{:2:}$

Fig. 3. The 3 first expression bases from the decomposition of the synthetic 3D data

Fig. 4. Sample data of the synthetic 3D dataset. Images 1 to 3 from the left show different identities and images 4 to 6 different expressions.

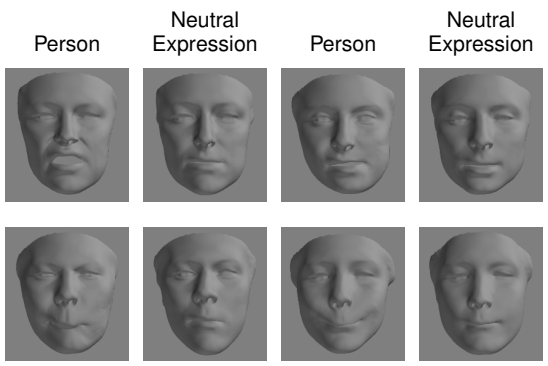


Fig. 5. Neutralising expressions

Fig. 6. Expression transfer

and 5 illuminations from the overall dataset. Our subset consists of 3675 images. We rigidly align the data to a mean shape in order to conserve the facial expression variations. Frequently, lighting becomes the first mode of variation in visual data.

We model illumination using first order spherical harmonics consisting of 4 components [41]. The decomposition can be adapted in this manner:

$$\boldsymbol{X} = \boldsymbol{B}_{(1)}(\boldsymbol{L} \odot \boldsymbol{E} \odot \boldsymbol{C}), \tag{23}$$

where $\boldsymbol{L} \in \mathbb{R}^{4\times n}$ is the matrix of first order spherical harmonic light coefficients, $\boldsymbol{E} \in \mathbb{R}^{e\times n}$ and $\boldsymbol{C} \in \mathbb{R}^{k\times n}$ represent expression and identity coefficients respectively.

Setting $e = 5$ and $k = 40$, we obtain a resulting tensor $\boldsymbol{\mathcal{B}} \in \mathbb{R}^{d\times 4\times e\times k}$. Our model indeed recovers illumination as the first mode of variation. The recovered basis $\boldsymbol{B}_{(1)}$, subject to orthogonality constraints, corresponds to a spherical harmonics basis and can be applied to estimate the normals and albedo of the object. The estimated normals are then warped back into the original space of the image and integrated using the method of [42] to recover the 3D reconstruction, see Figure 7.

Fig. 7. 3D Reconstruction on Multi-PIE [13] dataset

As the decomposition also decouples expression and identity variations into $\boldsymbol{E}$ and $\boldsymbol{C}$, we can use this to transfer facial expressions from one person to another person. Adapting the equation (4) to this decomposition (23), we specify for images $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ where the two images are of different people and expressions:

$$\boldsymbol{x}_i = \boldsymbol{\mathcal{B}} \times_2 \boldsymbol{l}_i \times_3 \boldsymbol{e}_i \times_4 \boldsymbol{c}_i, \quad \boldsymbol{x}_j = \boldsymbol{\mathcal{B}} \times_2 \boldsymbol{l}_j \times_3 \boldsymbol{e}_j \times_4 \boldsymbol{c}_j \tag{24}$$

By swapping $\boldsymbol{c}_i$ with $\boldsymbol{c}_j$, the identity coefficients, we can create a synthetic image $\boldsymbol{x}_{i_j}$ containing the expression of person i and identity of person j.

$$\boldsymbol{x}_{i_j} = \boldsymbol{\mathcal{B}} \times_2 \boldsymbol{l}_i \times_3 \boldsymbol{e}_i \times_4 \boldsymbol{c}_j. \tag{25}$$

In this way, a synthetic dataset of people with new expressions are created. Sample results of the expression transfer experiment are shown in Figure 8. Some of the examples are challenging ones such as transferring expressions across gender. The Multi-PIE [13] dataset contains a number of people wearing glasses which lead to artefacts in the area around the eyes in the synthetic images.

We test this synthetic data via an expression classification experiment to verify that the new synthetic expressions are recognisable. Specifically, we trained a linear SVM model with the original dataset and respective expression labels and used the synthetic dataset as test data. The prediction results are listed in Table 1. The high accuracy of 85.1% shows that the synthetic data manages to model the expressions contained in the original data.

5.3 Robust Disentanglement of Illumination and Shape

Using the robust decomposition method from Section 4.2, we show on two different dataset that the method is able to robustly reconstruct objects "in-the-wild". As "in-the-wild" data often contain noise and natural occlusions, a robust decomposition seems to be ideal in this case to separate the noise from the actual shape.

Fig. 8. Expression transfer on Multi-PIE. As our decomposition reduces the dimensionality of the images in the dataset, we show the images with the transferred expression next to the reconstructed image of the ground truth from the dataset. Given the decomposition, the reconstruction represents the result of a plausible expression transfer.

Data	Prediction accuracy
Synthetic expressions data	0.851

TABLE 1
Prediction accuracy on synthetic dataset

The decomposition applied in the below experiments is:

$$\boldsymbol{X} = \boldsymbol{B}_{(1)}(\boldsymbol{L} \odot \boldsymbol{C}) + \boldsymbol{E}, \tag{26}$$

where $\boldsymbol{L} \in \mathbb{R}^{4\times n}$ is the the matrix of first order spherical harmonic light coefficients, $\boldsymbol{C} \in \mathbb{R}^{k\times n}$ is a matrix of shape and identity coefficients and $\boldsymbol{E} \in \mathbb{R}^{d\times n}$ represents the matrix of sparse errors. A sparsity constraint has been put on $\boldsymbol{E}$

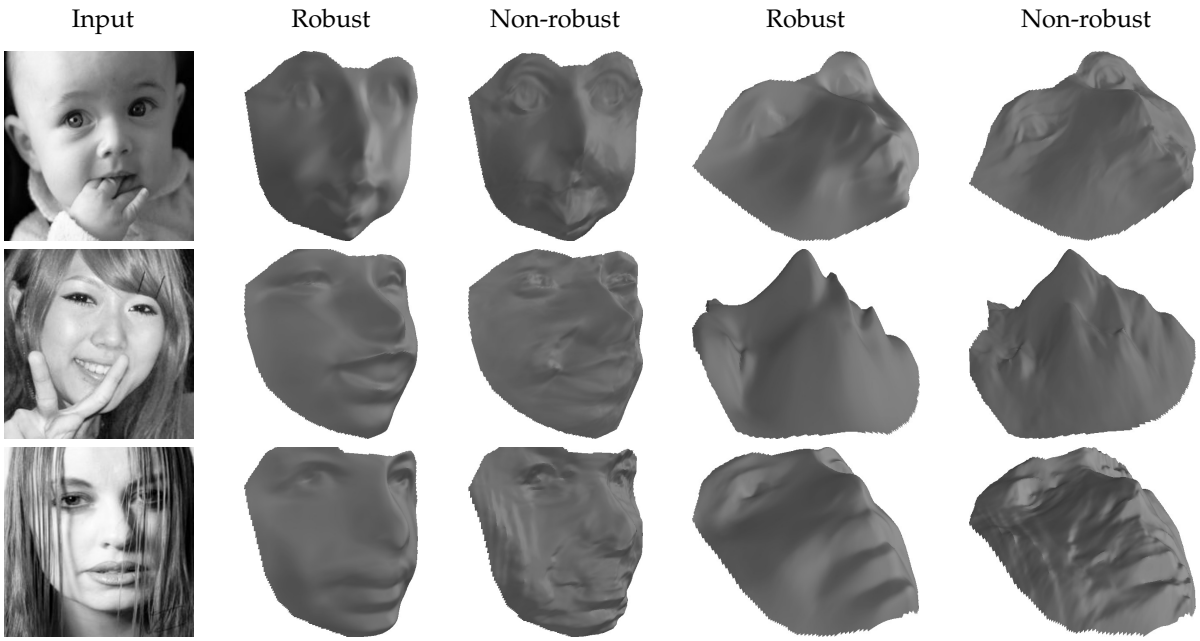


Fig. 9. **Comparison of the robust and non-robust decomposition**. Images from the HELEN [43] dataset.

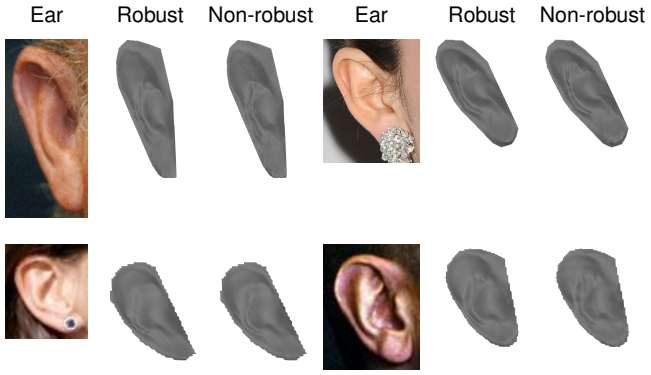


Fig. 10. Ear reconstructions. Images from the Ear dataset.

in the problem formulation. This is a special case of our proposed decomposition in (16).

In order to validate the usefulness of the robust decomposition, we have added 1% salt&pepper noise on the Photoface dataset. We then compare the estimated normals of our robust and non-robust decompositions on this noisy Photoface dataset. The "ground truth" normals were obtained from the clean Photoface dataset using PS [20]. Figure 12 shows the sample reconstructions from this experiment. From the 3D shape, we observe that the non-robust reconstruction also reconstructs the noise. The robust reconstruction obtains a smooth shape without the noise. The result of the quantitative evaluation can be found in Table 2. This clearly demonstrates that the robust decomposition outperforms the non-robust decomposition on noisy data.

We then show on two different "in-the-wild" datasets that the robust method outperforms the non-robust version.

Method	**Mean±Std**	**<30°**	**<35°**
Ours- Non-Robust	39.81°± 12.36°	0.7%	42.7%
Ours- Robust	**33.86°± 4.84°**	**16.4%**	**71.5%**

TABLE 2
Angular error for our method with and without robustness on Photoface containing 1% salt&pepper noise.

5.3.1 *Faces "In-the-wild"*

In this experiment, we show that our method is able to robustly reconstruct a large number of "in-the-wild" images. We use the HELEN [43] dataset containing 2000 identities with 1 image per person. We used the 68 facial landmarks from [44] for the warping to/from the mean reference shape. Figure 9 shows the results on a number of challenging images for $k = 200$. Clearly the robust method is able to separate illumination and appearance better than the non-robust method.

5.3.2 *Ears "In-the-wild"*

In this experiment, we show that our method works on other objects apart from faces. We collected 605 "in-the-wild" images of ears and annotated them with 55 landmarks. The landmarks were used for the warping to/from the mean reference shape. Setting $k = 100$, we apply our decomposition and show the results in Figure 10. The results indicate that the robust decomposition method outperforms the non-robust method.

5.4 Disentanglement of Illumination and Shape with Low-rank Constraints

Videos of a single person can specifically profit from the rank-constrained decomposition method from Section 4.3. As the rank-constrained decomposition also incorporates the l_1 norm, we can apply this method on noisy data.

We show this using two experiments: In the first experiment, we synthetically occlude part of a video with baboon patches. 20% of the frames in the video has been

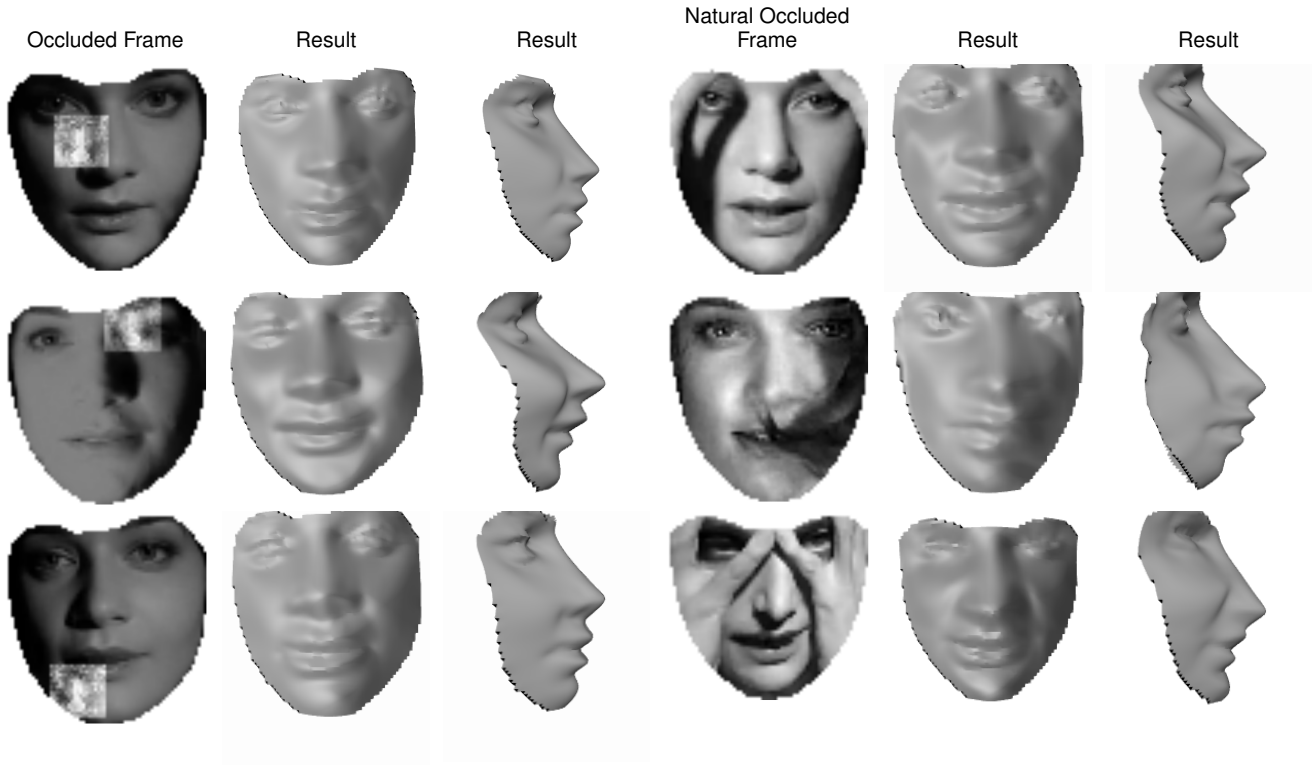


Fig. 11. Face reconstructions from occluded video frames using the rank-constrained decomposition.

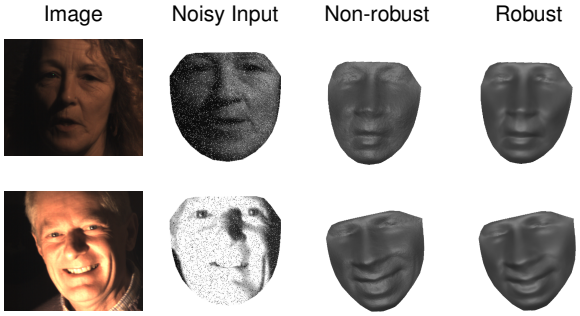


Fig. 12. Sample reconstruction from the Photoface dataset with 1% salt&pepper noise using non-robust and robust decomposition.

occluded by baboon patches. For each of those occluded frames, the baboon patch covers 10% of the frame. In the second experiment, we run the same method on videos where some frames have been naturally occluded by hands or hair.

Figure 11 shows the qualitative result of the two experiments. In both situations, we were able to reconstruct the faces quite well despite the occlusions. We can see how the reconstruction strongly resembles the person in the video.

Table 3 shows the quantitative result in the case of synthetic occlusion. The ground truth are the normals estimated without low-rank constraints on videos without occlusion ($k = 20$). We compare our methods with and without low-rank constraints ($k = 20$) on the videos containing occlusions. Clearly, our method with low-rank constraints is robust to occlusions and outperforms our method without low-rank constraints.

Method	**Mean±Std**	**<5°**	**<10°**
Ours-Without Low-rank Constraints	8.97° ± 2.02°	0%	88.5%
Ours-With Low-rank Constraints	**6.10° ± 1.74°**	**55.5%**	**100%**

TABLE 3
Angular error for our method with and without low-rank constraints on videos containing baboon patch occlusions.

5.5 Semi-Supervised Disentanglement of Expression and Identity

We have collected a new 3D database of people displaying 6 different expressions (happiness, disgust, anger, surprise, sadness and fear). We used NICP [45] to bring them in correspondence with the basel face model. In total we collect samples from 200 people, each sample was annotated with the expression label of the 6 different expressions. We applied the unsupervised version of the decomposition without graph-regularisation using the 6 labels. Keeping the identity parameters fixed to the ones of the mean face, we randomly sample values for $\boldsymbol{E}$. Figure 13 shows how we can generate 3D faces corresponding to each of the 6 expressions using this approach.

Then we split our dataset into 5 random splits, each time keeping 1000 samples for training and 200 for testing. We apply both the unsupervised and graph-regularised semi-supervised decomposition on the data and use a nearest neighbour classifier on $\boldsymbol{E}$ to predict the expressions of the test set. Table 4 shows how incorporating supervision via graph-regularisation strongly improves the expression classification accuracy.

Method	Accuracy
Unsupervised Decomposition	84.5%
Graph-regularised Semi-supervised Decomposition	**96.0%**

TABLE 4
Expression classification results using unsupervised and semi-supervised decomposition.

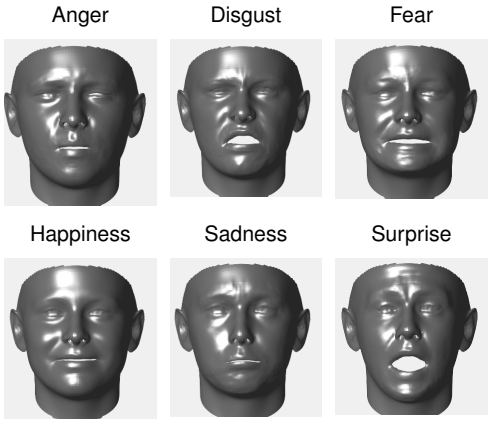


Fig. 13. 3D faces generated by keeping the identity component $\boldsymbol{C}$ fixed and randomly sampling the expression component $\boldsymbol{E}$.

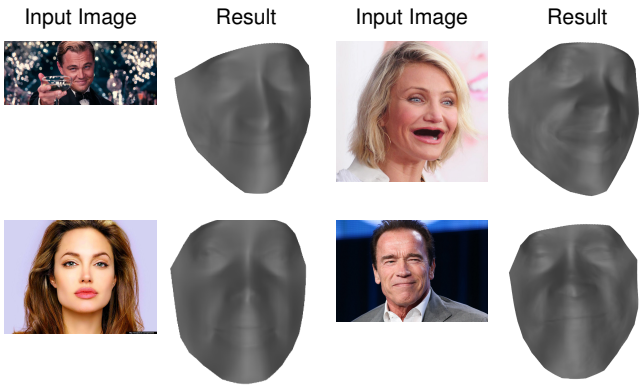


Fig. 14. Face reconstructions from single "in-the-wild" images using the deep unsupervised model trained on HELEN.

5.6 Unsupervised Normal Estimation using Deep Learning

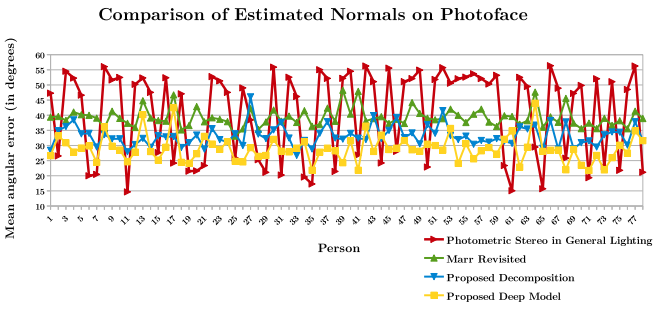


Fig. 15. Comparison of our two proposed methods with person-specific photometric stereo in general lighting of [21] and a generic state-of-the-art network [46]. The error has been calculated against the estimated normals from photometric stereo [20].

In this experiment, we use the normals estimated on the HELEN dataset [43] using the proposed method to train a 'fully convolutional' network to perform normal estimation on faces. This fully unsupervised pipeline consists of obtaining normals estimated using our decomposition and then feeding them as input data to a deep network. We use the network based on ResNet-50 [47] and the UberNet architecture [48]. UberNet used a Deep Convolutional Neural Network (DCNN) for surface normal estimation among a series of other tasks. The network architecture is in the supplementary material.

Method	Mean±Std against [20]	<35°	<40°
[21]	38.35° ± 15.63°	46.4%	46.8%
[46]	38.77° ± 3.27°	4.5%	73.0%
Ours-Decomposition Method	33.37° ± 3.29°	75.3%	96.3%
Ours-Unsupervised Deep Model	**28.53°** ± **4.23°**	**93.6%**	**97.8%**

TABLE 5
Angular error for the various surface normal estimation methods on the Photoface [49] dataset

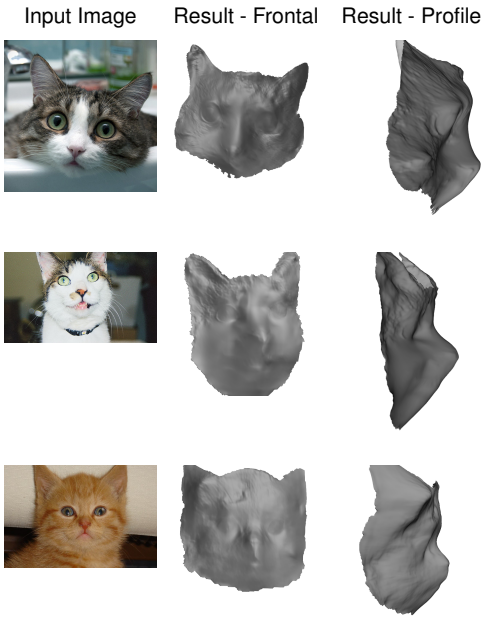


Fig. 16. Face reconstructions from single "in-the-wild" cats images using the deep unsupervised model trained on human faces and ears.

In order to quantitatively estimate the performance of the learned deep model, we require some level of ground truth data. Photoface [49] is a photometric stereo dataset containing single-view images of people taken under 4 different illumination conditions. We annotated 68 facial landmarks on 57 people from the dataset. The landmarks are used for the warping of the images into/from the mean reference shape for our proposed decomposition from Section 4.1. In the absence of ground truth depth or normal data, we use normals recovered from Photometric Stereo (PS) [20] as our ground truth. However, the normals from PS may be biased by outliers so these normals serve as a weak ground truth.

By testing our learned deep model on a previously unseen dataset such as Photoface is challenging as the images have been taken under different conditions to the ones in the HELEN training dataset. We plotted the mean angular error between our decomposition method (Section 4.1) results and the "ground truth" ones from PS [20] in Figure 15 and compare against our learned deep model.

Our decomposition method uses 2 randomly selected

images of a person under different lighting conditions. The method in [21] requires 4 images of a person under different lighting conditions. [46] is a generic state-of-the-art network which reconstructs from one image. Our deep model similarly only requires one image per person. From the quantitative results in Table 5, our deep model obtains a mean angular error of 28.53°across 273 people against 38.77°using [46]. It clearly shows that the deep model works very well on the Photoface test data and performs comparably and even slightly better than our decomposition method. The deep model obtains a mean angular error of 28.53°against 33.37°using the decomposition method. The reason for the strong performance is the variation of k. As the decomposition method is restricted by the number of annotated images in the Photoface dataset, the k used is 40. The deep model is trained on the larger HELEN dataset with $k = 400$. This suggests that the deep model may be able to extract more reconstruction details from the Photoface images than the decomposition method.

The results are extremely encouraging as they indicate that we can apply this unsupervised deep model directly to "in-the-wild" internet images of faces. Unlike the proposed decomposition method, the deep model does not require any warping of the images to/from the reference frame. This also is very beneficial for "in-the-wild" images. Figure 14 shows the reconstruction results of our deep model on internet images. The reconstructions nicely mirrors the facial expressions contained in the images.

In addition, we trained a separate network with the images from the HELEN and Ear datasets. The ground truth normals used were again the result of our decomposition method. We tested the model on human faces and found that it reconstructs more details. Then we tested this additionally on internet images of cats and found that due to the similar facial structure (eyes and nose), the model is able to reconstruct cat faces. The reconstructions can be seen in Figure 16. The model even manages to reconstruct the fur details, which is impressive.

6 Conclusion

We have proposed an unsupervised method able to discover the multilinear structure in visual data. To this end an alternating least squares algorithm has been developed. We extended this method to incorporate robustness and rank constraints. Our experiments show that the method is able to discover the multilinear structure of "in-the-wild" visual data without the presence of labels or well-organised input data. Additional experiments using an unsupervised deep learning pipeline show the application of the method directly on internet images of human as well as cat faces.

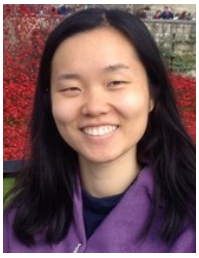
Acknowledgments

The work of Mengjiao Wang was funded by an EPSRC DTA from Imperial College London. The work of Dr. Zafeiriou was partially funded by EPSRC project FACER2VM.

References

[1] L. R. Fabrigar and D. T. Wegener, *Exploratory factor analysis*. Oxford University Press, 2011.
[2] H. Hotelling, "Analysis of a complex of statistical variables into principal components." *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.
[3] L. R. Tucker, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, vol. 31, no. 3, pp. 279–311, 1966.
[4] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM journal on Matrix Analysis and Applications*, vol. 21, no. 4, pp. 1253–1278, 2000.
[5] P. M. Kroonenberg and J. De Leeuw, "Principal component analysis of three-mode data by means of alternating least squares algorithms," *Psychometrika*, vol. 45, no. 1, pp. 69–97, 1980.
[6] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM review*, vol. 51, no. 3, pp. 455–500, 2009.
[7] J. B. Kruskal, "Rank, decomposition, and uniqueness for 3-way and n-way arrays," in *Multiway data analysis*. North-Holland Publishing Co., 1989, pp. 7–18.
[8] M. A. O. Vasilescu and D. Terzopoulos, "Multilinear analysis of image ensembles: Tensorfaces," in *European Conference on Computer Vision*. Springer, 2002, pp. 447–460.
[9] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 208–220, 2013.
[10] M. Signoretto, L. De Lathauwer, and J. A. Suykens, "Nuclear norms for tensors and their use for convex multilinear estimation," *Linear Algebra and Its Applications*, vol. 43, 2010.
[11] S. Gandy, B. Recht, and I. Yamada, "Tensor completion and low-n-rank tensor recovery via convex optimization," *Inverse Problems*, vol. 27, no. 2, p. 025010, 2011.
[12] T. Sim, S. Baker, and M. Bsat, "The cmu pose, illumination, and expression database," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1615 – 1618, December 2003.
[13] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.
[14] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3d facial expression database for facial behavior research," in *7th international conference on automatic face and gesture recognition (FGR06)*. IEEE, 2006, pp. 211–216.
[15] M. Wang, Y. Panagakis, P. Snape, and S. Zafeiriou, "Learning the multilinear structure of visual data," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
[16] I. Kemelmacher-Shlizerman, "Internet-based Morphable Model," *IEEE International Conference on Computer Vision (ICCV)*, 2013.
[17] P. Snape, Y. Panagakis, and S. Zafeiriou, "Automatic construction of robust spherical harmonic subspaces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 91–100.
[18] T. G. Kolda and B. W. Bader, "Tensor Decompositions and Applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, 2008.
[19] Q. Qiu and R. Chellappa, "Compositional dictionaries for domain adaptive face recognition," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5152–5165, 2015.
[20] R. J. Woodham, "Photometric method for determining surface orientation from multiple images," pp. 139–144, 1980.
[21] R. Basri, D. Jacobs, and I. Kemelmacher, "Photometric stereo with general, unknown lighting," *International Journal of Computer Vision*, vol. 72, no. 3, pp. 239–257, 2007.
[22] E. H. Adelson and A. P. Pentland, "The perception of shading and reflectance," *Perception as Bayesian inference*, pp. 409–423, 1996.
[23] R. Basri and D. W. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE T-PAMI*, vol. 25, no. 2, pp. 218–233, 2003.
[24] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: illumination cone models for face recognition under variable lighting and pose," *IEEE T-PAMI*, vol. 23, no. 6, pp. 643–660, 2001.
[25] R. Ramamoorthi and P. Hanrahan, "On the relationship between radiance and irradiance: determining the illumination from images of a convex lambertian object," *JOSA*, vol. 18, no. 10, pp. 2448–2459, 2001.
[26] R. T. Frankot and R. Chellappa, "A method for enforcing integrability in shape from shading algorithms," *IEEE T-PAMI*, vol. 10, no. 4, pp. 439–451, 1988.
[27] C. Khatri and C. R. Rao, "Solutions to some functional equations and their applications to characterization of probability distributions," *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 167–180, 1968.
[28] C. Bregler, A. Hertzmann, and H. Biermann, "Recovering non-rigid 3d shape from image streams," in *Computer Vision and Pattern*

Recognition, 2000. Proceedings. IEEE Conference on, vol. 2. IEEE, 2000, pp. 690–696.
[29] F. Roemer and M. Haardt, "Tensor-based channel estimation and iterative refinements for two-way relaying with multiple antennas and spatial reuse," *IEEE Transactions on Signal Processing*, vol. 58, no. 11, pp. 5720–5735, 2010.
[30] I. Kemelmacher-Shlizerman and S. M. Seitz, "Collection flow," in *CVPR*. IEEE, 2012, pp. 1792–1799.
[31] J. C. Gower and G. B. Dijksterhuis, *Procrustes problems*, ser. Oxford Statistical Science Series. Oxford, UK: Oxford University Press, January 2004, vol. 30.
[32] D. Bertsekas, "Constrained optimization and Lagrange multiplier methods," p. 410, 1982.
[33] R. Ramamoorthi, "Analytic PCA construction for theoretical analysis of lighting variability in images of a Lambertian object," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 10, pp. 1–12, 2002.
[34] S. Yan, D. Xu, B. Zhang, H. j. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40–51, Jan 2007.
[35] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using laplacianfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 328–340, March 2005.
[36] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.
[37] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ser. ICML'03. AAAI Press, 2003, pp. 912–919.
[38] A. M. Martinez and A. C. Kak, "Pca versus lda," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228–233, Feb 2001.
[39] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. Dunaway, "A 3D Morphable Model learnt from 10'000 faces," *Computer Vision and Pattern Recognition, IEEE Conference on (CVPR)*, pp. 5543–5552, 2016.
[40] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "FaceWarehouse: A 3D facial expression database for visual computing," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 3, pp. 413–425, 2014.
[41] R. Basri and D. Jacobs, "Lambertian reflectances and linear subspaces," *IEEE International Conference on Computer Vision*, vol. 00, no. C, pp. 383–390, 2001.
[42] R. T. Frankot and R. Chellappa, "Method for Enforcing Integrability in Shape from Shading Algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 10, no. 4, pp. 439–451, 1988.
[43] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7574 LNCS, no. PART 3, 2012, pp. 679–692.
[44] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: Database and results," *Image and Vision Computing*, vol. 47, pp. 3–18, 2016.
[45] S. Zafeiriou, A. Roussos, A. Ponniah, D. Dunaway, and J. Booth, "Large scale 3d morphable models," *International Journal of Computer Vision*, 2017.
[46] A. Bansal, B. C. Russell, and A. Gupta, "Marr revisited: 2d-3d alignment via surface normal prediction," *CVPR*, 2016.
[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *CVPR*, vol. 7, no. 3, pp. 171–180, 2016.
[48] I. Kokkinos, "Ubernet: Training a 'universal' convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory," *CoRR*, vol. abs/1609.02132, 2016.
[49] S. Zafeiriou, G. A. Atkinson, M. F. Hansen, W. A. P. Smith, V. Argyriou, M. Petrou, M. L. Smith, and L. N. Smith, "Face recognition and verification using photometric stereo: The photoface database and a comprehensive evaluation," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 1, pp. 121–135, 2013.



Mengjiao Wang is a PhD Student within the Department of Computing, Imperial College London, working under Dr. Stefanos Zafeiriou's supervision. Her research interests are in the areas of machine learning and computer vision, particularly disentangling meaningful variations from images. She received her MSc in Computational Statistics and Machine Learning, ordained with Distinction and Dean's List commendation, from University College London and her BSc in Mathematics and Computer Science, awarded with First Class Honours and Governor's Prize, from Imperial College London. Mengjiao received the scholarship of the Konrad Adenauer Foundation during her BSc and MSc studies.

Yannis Panagakis is a Lecturer (Assistant Professor equivalent) in Computer Science at Middlesex University London and a Research Fellow at the Department of Computing, Imperial College London. His research interests lie in machine learning and its interface with signal processing, high-dimensional statistics, and computational optimization. Specifically, Yannis is working on models and algorithms for robust and efficient learning from high-dimensional data and signals representing audio, visual, affective, and social information. He has been awarded the prestigious Marie-Curie Fellowship, among various scholarships and awards for his studies and research. Yannis currently serves as an Associate Editor of the Image and Vision Computing Journal. He co-organized the BMVC 2017 and several workshops and special sessions in top venues such as ICCV. He received his PhD and MSc degrees from the Department of Informatics, Aristotle University of Thessaloniki and his BSc degree in Informatics and Telecommunication from the University of Athens, Greece.

Patrick Snape received his PhD in Computer Science from Imperial College London, U.K. in 2017 under the supervision of Dr. Stefanos Zafeiriou. He was the recipient of the 2014 Qualcomm Innovation Fellowship. He has served as a reviewer for a number of computer vision conferences and journals including TPAMI, ICCV, CVPR and ECCV. His research focus is investigating the recovery of 3D shape from 2D images. His current research interests include Shape-from-Shading, deformable modelling, 3D image alignment and photometric stereo.

Stefanos P. Zafeiriou (M09) is currently a Reader in Machine Learning and Computer Vision with the Department of Computing, Imperial College London, London, U.K, and a Distinguishing Research Fellow with University of Oulu under Finish Distinguishing Professor Programme. He was a recipient of the Prestigious Junior Research Fellowships from Imperial College London in 2011 to start his own independent research group. He was the recipient of the President's Medal for Excellence in Research Supervision for 2016. He has received various awards during his doctoral and post-doctoral studies. He currently serves as an Associate Editor of the IEEE Transactions on Affective Computing and Computer Vision and Image Understanding journal. In the past he held editorship positions in IEEE Transactions on Cybernetics the Image and Vision Computing Journal. He has been a Guest Editor of over six journal special issues and co-organised over 13 workshops/special sessions on specialised computer vision topics in top venues, such as CVPR/FG/ICCV/ECCV (including three very successfully challenges run in ICCV13, ICCV15 and CVPR'17 on facial landmark localisation/tracking). He has co-authored over 55 journal papers mainly on novel statistical machine learning methodologies applied to computer vision problems, such as 2-D/3-D face analysis, deformable object fitting and tracking, shape from shading, and human behaviour analysis, published in the most prestigious journals in his field of research, such as the IEEE T-PAMI, the International Journal of Computer Vision, the IEEE T-IP, the IEEE T-NNLS, the IEEE T-VCG, and the IEEE T-IFS, and many papers in top conferences, such as CVPR, ICCV, ECCV, ICML. His students are frequent recipients of very prestigious and highly competitive fellowships, such as the Google Fellowship x2, the Intel Fellowship, and the Qualcomm Fellowship x3. He has more than 4500 citations to his work, h-index 36. He is the General Chair of BMVC 2017.