Learning the Multilinear Structure of Visual Data

Mengjiao Wang Yannis Panagakis Patrick Snape Stefanos Zafeiriou
Imperial College London
{m.wang15,i.panagakis,p.snape,s.zafeiriou}@imperial.ac.uk

Abstract

Statistical decomposition methods are of paramount importance in discovering the modes of variations of visual data. Probably the most prominent linear decomposition method is the Principal Component Analysis (PCA), which discovers a single mode of variation in the data. However, in practice, visual data exhibit several modes of variations. For instance, the appearance of faces varies in identity, expression, pose etc. To extract these modes of variations from visual data, several supervised methods, such as the TensorFaces, that rely on multilinear (tensor) decomposition (e.g., Higher Order SVD) 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 the first general multilinear method, to the best of our knowledge, that discovers the multilinear structure of visual data in unsupervised setting. That is, without the presence of labels. We demonstrate the applicability of the proposed method in two applications, namely Shape from Shading (SfS) and expression transfer.

1. Introduction

Statistical methods for data decomposition are cornerstones in statistics, image and signal processing, and computer vision. Probably, the most popular data decomposition method is the Principal Component Analysis (PCA) and the closely related Singular Value Decomposition (SVD).

Assuming that the data are stacked in the columns of a matrix, the PCA finds 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 therefore the SVD is unable to extract them. In order to disentangle the independent modes of variations, several multilinear (tensor) decomposition methods have been employed [25, 6, 15, 14, 16]. For example, the High Order SVD (HOSVD) [6] assume that the data are formed as a result of some multifactor confluence and aim at finding the different modes of variation by decomposing the a carefully designed data tensor. Having found the HOSVD, the data admit to a linear analysis per mode, since each mode is allowed to vary in turn, while the remaining modes are held constant. For instance, assume a population of faces with differing identities, expressions and seen under different views (poses), then the HOSVD decomposes the population into different modes of variation for expressions, identities, and poses per pixel. This method is known as TensorFaces [26]. Thus, having disentangled the modes of variation, we can vary expressions independently to identity and pose.

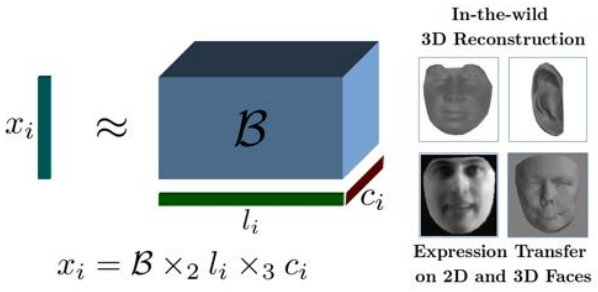

Figure 1: Visualisation of the unsupervised multilinear decomposition and its applications. A sample vector $\mathbf{x}_i$ is assumed to be generated by a common multilinear structure $\mathbf{B}$ (in this example with two modes) and sample specific weights $\mathbf{l}_i$ and $\mathbf{c}_i$.

The main limitation in applying the above multilinear decompositions is that they require a complete data tensor which has to be built using labels for each mode of variation. That is, each mode of variation must be represented in data. Using the aforementioned example of faces with varying expression, identity and pose, we require that for each and every person samples for every possible expression and pose should exist in order to build the required

complete tensor[1]. Clearly, these requirements limit the applicability of multilinear decompositions to data captured in controlled conditions (e.g. PIE [23], Multi-PIE [10] and BU-3DFE [28]), where it can be guaranteed that all the necessary data variations along with their labels are available.

This paper is concerned with the problem of **unsupervised multilinear decomposition** of data, a problem which, even though it has many applications, has received very limited attention. In particular, we propose the first, to the best of our knowledge, multilinear decomposition which finds the potential multilinear structure of data without labels or requiring a complete set of data. That is, assuming a set of vectorial data stacked in a matrix, our aim is find the multilinear structure and the corresponding weights (coefficients) that best explain the data under a known number of variations. The proposed model is schematically summarised in Figure 1. We show that the proposed methodology indeed disentangles the modes of variations without any labels, as well as having a wide number of applications including 3D object reconstruction "in-the-wild".

The contributions of the paper are summarised as follows:

- We propose a novel unsupervised decomposition that recovers the multilinear structure of visual data and thus an arbitrary number of different modes of variation.
- We develop an efficient alternating least squares type of algorithm to perform the decomposition.
- We highlight the relationship between the proposed decomposition and recent Shape from Shading (SfS) methods [11, 24], and show that the methods in [11, 24] are very special cases of the proposed general multilinear decomposition.
- We demonstrate the usefulness of the proposed decomposition to various applications including SfS and facial expression transfer.

2. Related Work

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 [23] and Multi-PIE [10] 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

[1] Methods for completing the tensor have been proposed but they are only approximate[18, 22, 8].

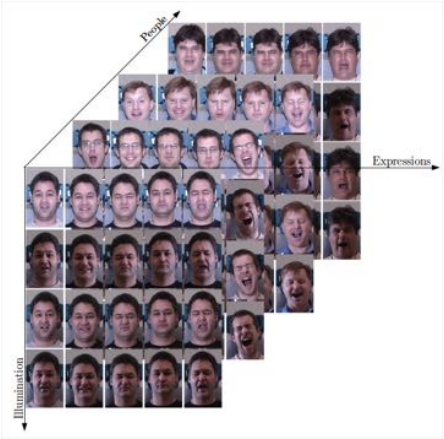

Figure 2: Visualisation of the Multi-PIE [10] dataset. Collecting data where every person is present in all the lighting and expression variations is an expensive process that does not scale well.

motivated the use of multilinear decompositions, such as HOSVD [6], in order to disentangle the different modes of variations, with the most popular being the TensorFaces representation [26]. Nevertheless, even though the methods 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 tensor decompositions [19] are still mainly applied to tightly controlled databases such as PIE and Multi-PIE, visualised in Figure 2, and not to "in-the-wild" data.

A seemingly unrelated area of research that relies heavily on data decomposition is that of facial SfS [27] and Uncalibrated Photometric Stereo in General Lighting (UPS) [2]. Starting from a totally different perspective, the current state-of-the-art object-specific UPS techniques [11, 24] perform a rank constrained Khatri-Rao (KR) factorization [12]. The first paper where the decomposition has been proposed and applied in 3D facial shape reconstruction was [11]. [11] was inspired by the decomposition techniques employed in the related area of Structure-from-Motion [4]. Recently, [24] made the link between the KR factorisation and the UPS. In this paper, we show that the decompositions proposed in [11, 24] are very special cases of the proposed unsupervised tensor decomposition. Furthermore, the proposed decomposition goes much further than the special case [11, 24] and can be used for disentangling an arbitrary number of modes of variation.

3. 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., $\boldsymbol{\mathcal{X}}$. The *order* of a tensor is the number of indices needed to address its elements. Consequently, each element of an Mth-order tensor $\boldsymbol{\mathcal{X}}$ is addressed by M indices, i.e., $(\boldsymbol{\mathcal{X}})_{i_1,i_2,\ldots,i_M} \doteq x_{i_1,i_2,\ldots,i_M}$.

The sets of real and integers numbers is denoted by $\mathbb{R}$ and $\mathbb{Z}$, respectively. A set of N 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 $\boldsymbol{\mathcal{X}}$ 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 $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_M}$ 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, $\boldsymbol{\mathcal{X}} = \mathbf{x}^{(1)} \circ \mathbf{x}^{(2)} \circ \cdots \circ \mathbf{x}^{(M)} \doteq \bigcirc_{m=1}^{M} \mathbf{x}^{(m)}$, where $\circ$ denotes for the vector outer product.

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

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

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

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

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 [13] for example.

Finally, $\|\cdot\|_F$ denotes the Frobenius norm.

4. Proposed Method

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:

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

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., [13], the above decomposition is written in matrix form as

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

where $\boldsymbol{B_{(1)}} \mathbb{R}^{d \times K_2 \cdot K_3 \cdots K_M}$ 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 [25] and the HOSVD [6].

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

$$\begin{aligned} \underset{\boldsymbol{B_{(1)}}, \{\boldsymbol{A}^{(m)}\}_{m=2}^{M}}{\arg\min} \; & \|\boldsymbol{X} - \boldsymbol{B_{(1)}}\big(\bigodot_{m=2}^{M} \mathbf{A}^{(m)}\big)\|_F^2 \\ \text{s.t. } & \boldsymbol{B}_{\boldsymbol{(1)}}^T \boldsymbol{B_{(1)}} = \mathbf{I}. \end{aligned} \tag{4}$$

Optimisation problem (4) is non-convex. Therefore, we propose to solve (4) 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:

$$\begin{aligned} \boldsymbol{B_{(1)}}[t+1] = \underset{\boldsymbol{B_{(1)}}}{\arg\min} \; & \|\boldsymbol{X} - \boldsymbol{B_{(1)}}\big(\bigodot_{m=2}^{M} \mathbf{A}^{(m)}[t]\big)\|_F^2 \\ \text{s.t. } & \boldsymbol{B}_{\boldsymbol{(1)}}^T \boldsymbol{B_{(1)}} = \mathbf{I}. \end{aligned} \tag{5}$$

$$\{\boldsymbol{A}^{(m)}[t+1]\}_{m=2}^{M} = \underset{\{\boldsymbol{A}^{(m)}\}_{m=2}^{M}}{\arg\min} \; \|\boldsymbol{X} - \boldsymbol{B_{(1)}}[t+1]\big(\bigodot_{m=2}^{M} \mathbf{A}^{(m)}\big)\|_F^2 \tag{6}$$

Solving (5): Problem (5) is an orthogonal Procrustes problem, whose solution is given by [9]: $\boldsymbol{B_{(1)}}[t+1] = \boldsymbol{U}\boldsymbol{V}^T$, 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 (6): Due to the unitary invariance of the Frobenius norm (5) is equivalent to

$$\underset{\{\boldsymbol{A}^{(m)}\}_{m=2}^{M}}{\arg\min} \; \|\boldsymbol{B_{(1)}}[t+1]^T \boldsymbol{X} - \bigodot_{m=2}^{M} \mathbf{A}^{(m)}\|_F^2, \tag{7}$$

which is a Khatri-Rao factorisation problem [20]. 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:

$$\boldsymbol{q}_i = \bigodot_{m=2}^{M} \mathbf{a}_i^{(m)} \tag{8}$$

Let us partition $\boldsymbol{q}_i$ into a set $K_{M-1} \cdot K_{M-2} \cdots K_2$ 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

$\boldsymbol{q}_i = [\boldsymbol{q}_i^{B_1{}^T} \boldsymbol{q}_i^{B_2{}^T} \cdots \boldsymbol{q}_i^{B_{K_{M-1}\cdot K_{M-2}\cdots K_2}{}^T}]^T$. This partitioning enables us to rearranging the elements of $\boldsymbol{q}_i$ into a tensor $\boldsymbol{\mathcal{Q}}_i \in \mathbb{R}^{K_M \times K_{M-1} \times \cdots \times K_2}$ 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 (8), $\boldsymbol{\mathcal{Q}}_i$ is written as

$$\boldsymbol{\mathcal{Q}}_i = \mathbf{a}_i^{(M)} \circ \mathbf{a}_i^{(M-1)} \circ \cdots \circ \mathbf{a}_i^{(2)} \tag{9}$$

Equation (9) 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 $\boldsymbol{\mathcal{Q}}_i$, for $i = 1, 2, \ldots, N$. An efficient way to find the best rank-1 approximation of $\boldsymbol{\mathcal{Q}}_i$ is to exploit the truncated HOSVD [6]. That is,

$$\boldsymbol{\mathcal{Q}}_i = s \prod_{n=1}^{M-1} \times_n \mathbf{u}_i^{(n)}, \tag{10}$$

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 $\boldsymbol{\mathcal{Q}}_i$ 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

$$\boldsymbol{a}_i^{(m)} = s^{\frac{1}{M-1}} \mathbf{u}_i^{(M-m+1)}, \tag{11}$$

for $m = 2, 3, \ldots, M$. Interestingly, the estimation of the variation coefficients according to (11) resolves the inherent scaling ambiguity in (7) by assigning the same Euclidean-norm to each column of $\mathbf{A}^{(m)}$. The procedure of solving (4) is summarised in Algorithm 1.

Remarks: In the special case of 2 modes and where $K_2 = 4$, (3) becomes:

$$\boldsymbol{X} = \boldsymbol{B_{(1)}}(\boldsymbol{L} \odot \boldsymbol{C}), \tag{12}$$

where $\boldsymbol{L} = \boldsymbol{A}_2 \in \mathbb{R}^{4\times n}$, $\boldsymbol{C} = \boldsymbol{A}_3 \in \mathbb{R}^{k\times n}$. Equation (12) corresponds to the formulation used by [24][2].

Let $\boldsymbol{P} = \boldsymbol{L} \odot \boldsymbol{C}$ then,

$$\boldsymbol{X} = \boldsymbol{B_{(1)}}\boldsymbol{P}. \tag{13}$$

Equation (13) corresponds to the formulation used by [11]. $\boldsymbol{P} = \boldsymbol{L} \odot \boldsymbol{C}$ has been implied by [11] but not explicitly formulated as such. Hence this shows that [11, 24] represent special cases of our general decomposition.

[2] A minor difference in [24] is the separation of $\boldsymbol{X}$ into a low-rank $\boldsymbol{A}$ and sparse error $\boldsymbol{E}$

Algorithm 1: Multilinear Tensor Decomposition Algorithm

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

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 *each image* $i = 1 \ldots N$ **do**
    construct $\boldsymbol{\mathcal{Q}}_i \in \mathbb{R}^{K_M \times K_{M-1} \cdots \times K_2}$ from $\boldsymbol{q}_i[t]$
    $[\boldsymbol{S}_i, \boldsymbol{U}_i] \leftarrow$HOSVD$(\boldsymbol{\mathcal{Q}}_i)$
    for *each mode m = 2 ... M* **do**
      $\boldsymbol{a}_i^{(m)}[t] = (\boldsymbol{S}_i)_1^{\frac{1}{M-1}} \boldsymbol{U}_i^{(M-m+1)}$
    end
  end
  $[\boldsymbol{U}, \boldsymbol{\Sigma}, \boldsymbol{V}] \leftarrow$SVD$(\boldsymbol{X}\big(\bigodot_{m=2}^M \mathbf{A}^{(m)}[t]\big)^T)$
  $\boldsymbol{B_{(1)}}[t+1] = \boldsymbol{U}\boldsymbol{V}^T$
  $\boldsymbol{Q}[t+1] = \boldsymbol{B}_{(1)}[t+1]^T\boldsymbol{X}$
  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
Tensorise $\boldsymbol{B_{(1)}}$ into $\boldsymbol{\mathcal{B}} \in \mathbb{R}^{d\times K_2 \times \cdots \times K_M}$

5. Experiments

In this section we provide a number of experiments to show that our model recovers meaningful modes of variations. All the data used have been aligned to a reference shape to achieve pixelwise correspondence.

Frequently the first mode of variation in visual data is lighting. Hence we first consider data containing lighting variations i.e., objects under different lights and decompose the data into illumination and shape/identity components. 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 this decomposition can be applied to in-the-wild datasets of different objects.

We then investigate synthetic 3D facial data that contains both facial expression and identity variations. We show that our decomposition correctly decouples expression and identity.

Finally, we extend the decomposition to data that simultaneously contains lighting, expression and identity variations.

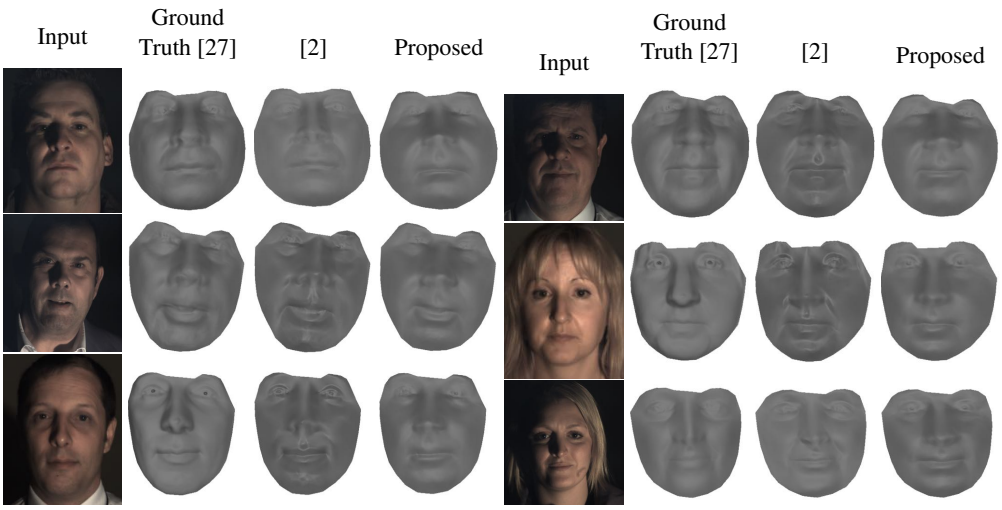

Figure 3: 3D shape reconstruction: Comparison of our proposed method with photometric stereo [27] and the person-specific photometric stereo in general lighting of [2]. Images from the Photoface [29] dataset.

5.1. Illumination and Shape Decomposition

Given a dataset of objects in piecewise correspondence (*e.g.* warped to a mean reference shape) but containing light and identity variations, we seek to recover the illumination mode of variation. We model illumination using first order spherical harmonics consisting of 4 components [1]:

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

where $\boldsymbol{B}_{(\mathbf{1})} \in \mathbb{R}^{d \times 4k}$ is the orthogonal mode-1 matricisation of our proposed tensor $\boldsymbol{\mathcal{B}}$, $\boldsymbol{L} \in \mathbb{R}^{4 \times n}$ is the matrix of first order spherical harmonic light coefficients and $\boldsymbol{C} \in \mathbb{R}^{k \times n}$ is a matrix of shape and identity coefficients. Evidently, this is a special case of our proposed decomposition in (3). The choice of k is subject to a trade-off between reconstruction detail of the images and the ability of the decomposition to separate illumination and shape/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 lighting from shape and identity. Our model indeed recovers illumination as the first mode of variation. The recovered basis $\boldsymbol{B}_{(\mathbf{1})}$, 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 [7] to recover the 3D reconstruction. We run this experiment on a variety of a number datasets including Photoface [29], HELEN [17] and a collected set of human ear images.

5.1.1 Comparison using Photoface

Photoface [29] is a photometric stereo dataset containing single-view images of people taken under 4 different illumination conditions. We annotated 68 facial landmarks on 273 people from the dataset. The landmarks are used for the warping of the images into/from the mean reference shape. In the absence of ground truth depth or normal data, we use normals recovered from Photometric Stereo (PS) [27] as our ground truth. However, the normals from PS may be biased by outliers so these normals serve as a weak ground truth.

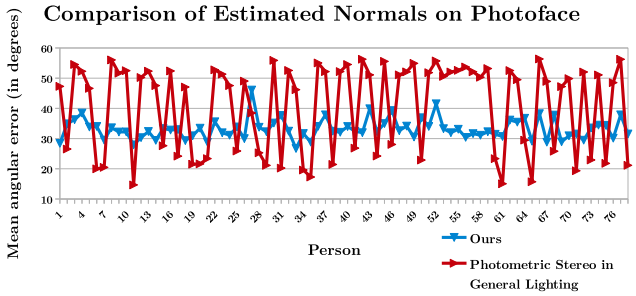

Figure 4: Comparison of our proposed method with person-specific photometric stereo in general lighting of [2]. The error has been calculated against the estimated normals from photometric stereo [27].

We wish to show that our decomposition works even in the case of an incomplete tensor. To this end, we apply our algorithm to a subset of the dataset: for each person we randomly choose 2 out of the 4 images. The data is incomplete as we do not have the same set of images for each person. For this experiment we set $k = 40$.

We compare our results against the person-specific results of [2] which utilises all 4 lighting conditions and applied the method per person. Figure 3 shows the sample

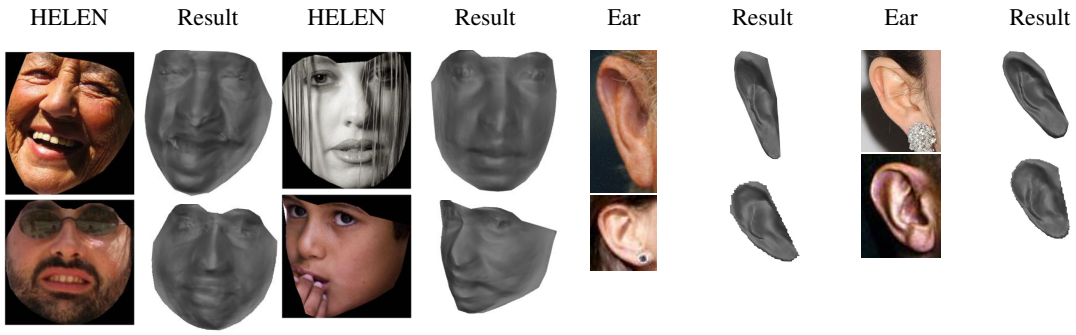

Figure 5: Face and ear reconstructions. Images from the HELEN [17] and Ear datasets.

Method	Mean angular error against [27]
[2]	$38.35° \pm 15.63°$
Ours	$\mathbf{33.37°} \pm \mathbf{3.29°}$

Table 1: Comparison of estimated normals

reconstructions from this experiment. We also plotted the mean angular error between our results and the "ground truth" ones from PS [27] in Figure 4 and compare against [2]. We can see that even with missing light information and across multiple identities, our model achieves competitive results, see quantitative results in Table 1. We obtain a mean angular error of $33.27°$across all 273 people against $38.35°$using [2]. In addition our method tends to more robust with $\pm 3.29°$ of standard deviation compared with $\pm 15.63°$ from [2]. These results are the only quantitative results we can obtain as the other datasets do not provide the necessary light information to compute "ground truth" normals from PS.

5.1.2 Comparison using "in-the-wild" Datasets

In this experiment, we show that our method is able to reconstruct a large number of in-the-wild images. In the first experiment, we use the HELEN [17] dataset containing 2000 identities with 1 image per person. We used the 68 facial landmarks from [21] for the warping to/from the mean reference shape. Figure 5 shows the results on a number of challenging images for $k = 40$. [2] cannot be run on this dataset as it would require at least 4 different lightings per identity.

In the second experiment, we show that our method works on other objects apart from faces. We apply the same methodology to in-the-wild images of ears. We used the 605 images of ears annotated with 55 landmarks of [30]. Setting $k = 75$, we apply our decomposition and show the results in Figure 5.

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

5.2. Expression and Identity Decomposition

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 [3] and put in correspondence with the blendshapes of the FaceWarehouse [5] so that we can allow for expressions. We used 200 components to describe identity and 9 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 6.

The decomposition becomes:

$$\boldsymbol{X} = \boldsymbol{B}_{(\mathbf{1})}(\boldsymbol{E} \odot \boldsymbol{C}), \tag{15}$$

where $\boldsymbol{E} \in \mathbb{R}^{e \times n}$ 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.

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 2 bases are plotted in Figure 7. We also discover that the first basis $\pm\boldsymbol{\mathcal{B}}_{:0:}$, visualised in Figure 7a, is a basis of neutral expressions. This is impressive as the neutral expression did not exist in the original dataset.

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

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

Figure 7: The 2 first expression bases from the decomposition of the synthetic 3D data

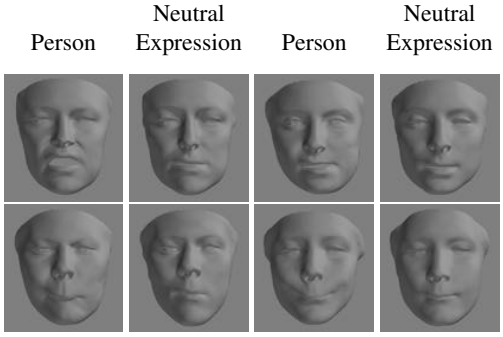

Figure 8: Neutralising expressions

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 $\boldsymbol{\mathcal{B}}_{:0:}$:

$$\boldsymbol{x}'_i = \boldsymbol{B}_0(\boldsymbol{B}_0^T\boldsymbol{B}_0)^{-1}\boldsymbol{B}_0^T\boldsymbol{x}_i, \tag{16}$$

where $\boldsymbol{x}'_i$ denotes the resulting neutral face of the person in $\boldsymbol{x}_i$. The results are visualised in Figure 8.

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 9.

Comparing $\boldsymbol{E}\boldsymbol{E}^T$ obtained using our unsupervised method to the $\boldsymbol{U}_{exp}$ obtained by the supervised TensorFaces [26], we find that components of $\boldsymbol{E}\boldsymbol{E}^T$ can achieve correlations of 0.66 with $\boldsymbol{U}_{exp}$.

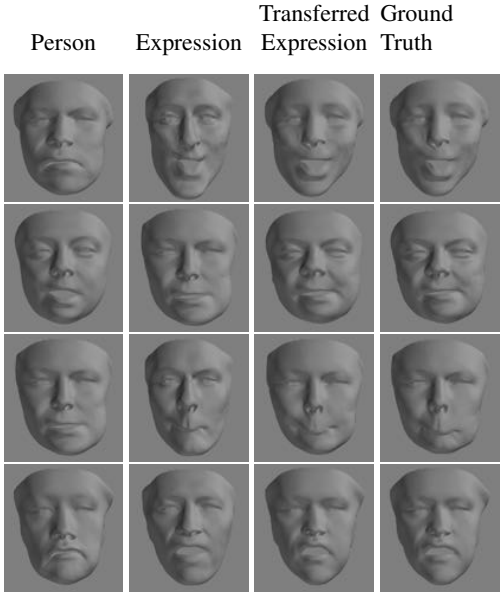

Figure 9: Expression transfer

5.3. Illumination, Expression and Identity Decomposition

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 [10] dataset. We select 147 identities, 5 expressions 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.

The decomposition can be adapted in this manner:

$$\boldsymbol{X} = \boldsymbol{B}_{(\mathbf{1})}(\boldsymbol{L} \odot \boldsymbol{E} \odot \boldsymbol{C}), \tag{17}$$

where $\boldsymbol{L} \in \mathbb{R}^{4\times n}$, $\boldsymbol{E} \in \mathbb{R}^{e\times n}$ and $\boldsymbol{C} \in \mathbb{R}^{k\times n}$ represent lighting, 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 experiments show that lighting remains the first mode of variation and we are still able to reconstruct 3D shape from this information, see Figure 10.

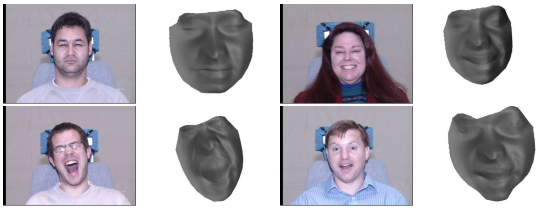

Figure 10: 3D Reconstruction on Multi-PIE [10] dataset

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 (2) to this decomposition (17), we specify for images $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ where the two images are of different people and expressions:

$$\boldsymbol{x}_i = \boldsymbol{\mathcal{B}} \times_2 \boldsymbol{l}_i \times_3 \boldsymbol{e}_i \times_4 \boldsymbol{c}_i, \quad \boldsymbol{x}_j = \boldsymbol{\mathcal{B}} \times_2 \boldsymbol{l}_j \times_3 \boldsymbol{e}_j \times_4 \boldsymbol{c}_j \tag{18}$$

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.

$$\boldsymbol{x}_{i_j} = \boldsymbol{\mathcal{B}} \times_2 \boldsymbol{l}_i \times_3 \boldsymbol{e}_i \times_4 \boldsymbol{c}_j. \tag{19}$$

In this way, a synthetic dataset of people with new expressions are created. Sample results of the expression transfer experiment are shown in Figure 11. Some of the examples are challenging ones such as transferring expressions across gender. The Multi-PIE [10] dataset contains a number of people wearing glasses which lead to artefacts in the area around the eyes in the synthetic images. 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.

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 2. The high accuracy of 85.1% shows that the synthetic data manages to model the expressions contained in the original data.

Figure 11: Expression transfer on Multi-PIE

Data	Prediction accuracy
Synthetic expressions data	0.851

Table 2: Prediction accuracy on synthetic dataset

6. Conclusion

We have proposed a unsupervised method able to discover the multilinear structure in visual data. To this end an alternating least squares algorithm has been developed. 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.

Acknowledgements

Mengjiao Wang is funded by an EPSRC DTA from Imperial College London. Stefanos Zafeiriou is partially supported by the EPSRC project EP/N007743/1 (FACER2VM).

References

[1] R. Basri and D. Jacobs. Lambertian reflectances and linear subspaces. *IEEE International Conference on Computer Vision*, 00(C):383–390, 2001.

[2] R. Basri, D. Jacobs, and I. Kemelmacher. Photometric stereo with general, unknown lighting. *International Journal of Computer Vision*, 72(3):239–257, 2007.

[3] 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)*, pages 5543–5552, 2016.

[4] 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*, volume 2, pages 690–696. IEEE, 2000.

[5] 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*, 20(3):413–425, 2014.

[6] L. De Lathauwer, B. De Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000.

[7] R. T. Frankot and R. Chellappa. Method for Enforcing Integrability in Shape from Shading Algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(4):439–451, 1988.

[8] S. Gandy, B. Recht, and I. Yamada. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems*, 27(2):025010, 2011.

[9] J. C. Gower and G. B. Dijksterhuis. *Procrustes problems*, volume 30 of *Oxford Statistical Science Series*. Oxford University Press, Oxford, UK, January 2004.

[10] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-PIE. *Image and Vision Computing*, 28(5):807–813, 2010.

[11] I. Kemelmacher-Shlizerman. Internet-based Morphable Model. *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[12] 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*, pages 167–180, 1968.

[13] T. G. Kolda and B. W. Bader. Tensor Decompositions and Applications. *SIAM Review*, 51(3):455–500, 2008.

[14] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[15] P. M. Kroonenberg and J. De Leeuw. Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika*, 45(1):69–97, 1980.

[16] J. B. Kruskal. Rank, decomposition, and uniqueness for 3-way and n-way arrays. In *Multiway data analysis*, pages 7–18. North-Holland Publishing Co., 1989.

[17] 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)*, volume 7574 LNCS, pages 679–692, 2012.

[18] 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*, 35(1):208–220, 2013.

[19] Q. Qiu and R. Chellappa. Compositional dictionaries for domain adaptive face recognition. *IEEE Transactions on Image Processing*, 24(12):5152–5165, 2015.

[20] 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*, 58(11):5720–5735, 2010.

[21] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing*, 47:3–18, 2016.

[22] 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*, 43, 2010.

[23] T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression database. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1615 – 1618, December 2003.

[24] 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*, pages 91–100, 2015.

[25] L. R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

[26] M. A. O. Vasilescu and D. Terzopoulos. Multilinear analysis of image ensembles: Tensorfaces. In *European Conference on Computer Vision*, pages 447–460. Springer, 2002.

[27] R. J. Woodham. Photometric method for determining surface orientation from multiple images, 1980.

[28] 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)*, pages 211–216. IEEE, 2006.

[29] 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*, 8(1):121–135, 2013.

[30] Y. Zhou and S. Zafeiriou. Deformable models of ears in-the-wild for alignment and recognition. In *FG'2017*, June 2017.