Imperial College London

IEEE 2017 Conference on Computer Vision and Pattern Recognition

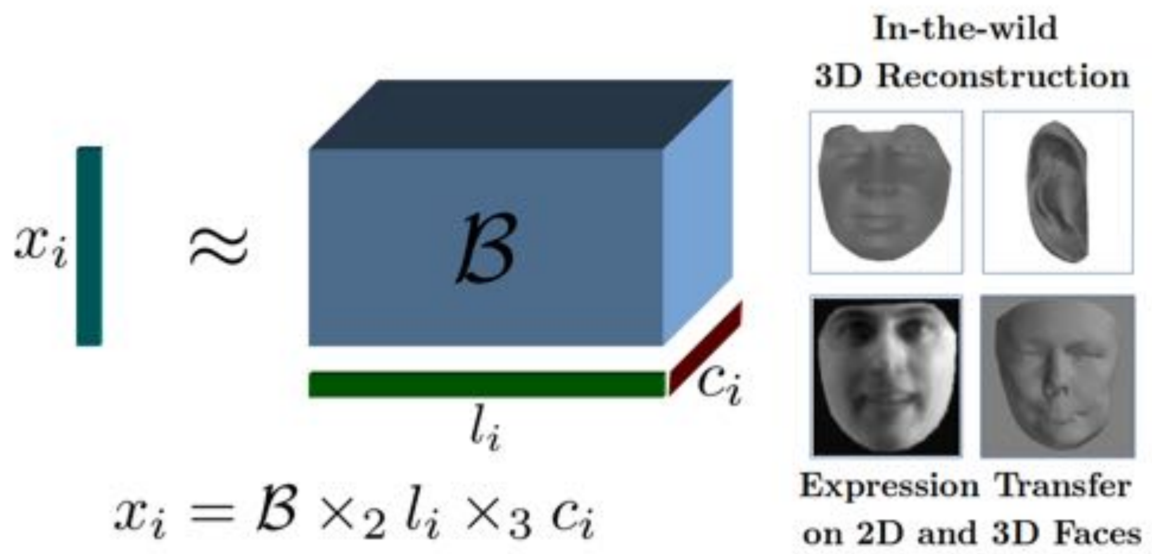
CVPR July 21-26 HONOLULU 2017

Learning the Multilinear Structure of Visual Data

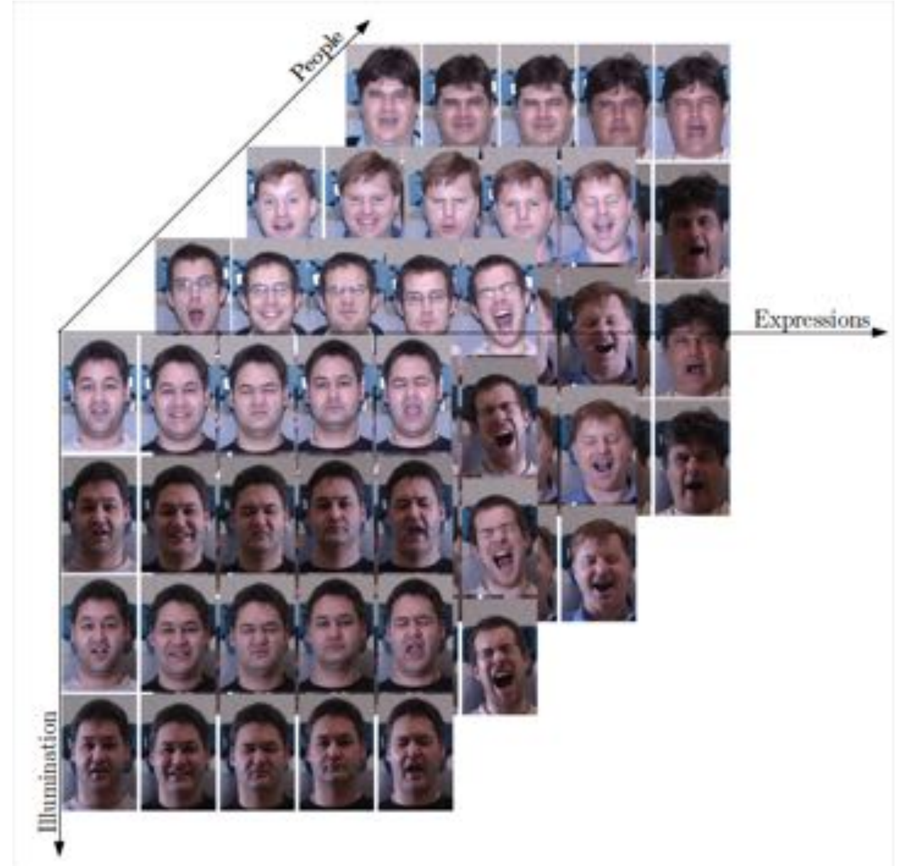
Mengjiao Wang, Yannis Panagakis, Patrick Snape and Stefanos Zafeiriou
Imperial College London, UK

1. Introduction

We propose the **first** general multilinear method that discovers the **multilinear structure of visual data** in **unsupervised setting** (i.e., without the presence of labels).



2. Motivation



- Statistical decomposition methods are important for discovering the modes of variations of visual data.
- Methods such as Principal Component Analysis (PCA) discovers a **single** mode of variation in the data. In practice, visual data exhibit several modes of variations e.g. identity, expression, pose etc.
- Multilinear decomposition methods developed are **supervised** such as TensorFaces [4]. 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**.

3. Prior Work

- Class-specific uncalibrated photometric stereo techniques [2, 3] perform a rank constrained Khatri-Rao (KR) factorization to reconstruct 3D shapes. The decompositions in [2, 3] are special cases of our method.
- Our unsupervised tensor decomposition method can be used for **disentangling an arbitrary number of modes of variation**.

4. Proposed Method

Model

We propose the following decomposition to discover $M-1$ different modes of variation

$$\boldsymbol{x}_i = \boldsymbol{\mathcal{B}} \times_2 \boldsymbol{a}_i^{(2)} \times_3 \boldsymbol{a}_i^{(3)} \cdots \times_m \boldsymbol{a}_i^{M} \quad (1)$$

$\boldsymbol{\mathcal{B}} \in \mathbb{R}^{d \times K_2 \times \cdots \times K_M}$: common multilinear basis of $\boldsymbol{X} \in \mathbb{R}^{d \times N}$, matrix of observations.
$\{\boldsymbol{a}_i^{(m)} \in \mathbb{R}^{K_m}\}_{m=2}^{M}$: variation coefficients in each mode specific to the image $\boldsymbol{x}_i$.
(1) in matrix form is

$$\boldsymbol{X} = \boldsymbol{B}_{(1)}(\boldsymbol{A}^{(2)} \odot \boldsymbol{A}^{(3)} \cdots \odot \boldsymbol{A}^{(M)}) \quad (2)$$

$\boldsymbol{B}_{(1)} \in \mathbb{R}^{d \times K_2 \cdot K_3 \cdot K_M}$: the mode-1 matricisation of $\boldsymbol{\mathcal{B}}$.
$\odot$ denotes the Khatri-Rao (column-wise Kronecker product) product of matrices.

Optimisation

We solve

$$\begin{aligned} &\underset{\boldsymbol{B}_{(1)}, \{\boldsymbol{A}^{(m)}\}_{m=2}^{M}}{\arg\min} \|\boldsymbol{X} - \boldsymbol{B}_{(1)}(\bigodot_{m=2}^{M} \mathbf{A}^{(m)})\|_F^2 \\ &\text{s.t. } \boldsymbol{B}_{(1)}^T \boldsymbol{B}_{(1)} = \mathbf{I}. \end{aligned} \quad (3)$$

by employing Alternating Least Squares (ALS).

5. Shape and Illumination

$$\boldsymbol{X} = \boldsymbol{B}_{(1)}(\boldsymbol{L} \odot \boldsymbol{C}) \quad (4)$$

Input Data: Different people, 4 different illuminations, single view

Result: We are able to disentangle shape from illumination and use this to reconstruct the 3D shape of the person. Our method outperforms [1] even on incomplete data (2 vs 4 illumination per person).

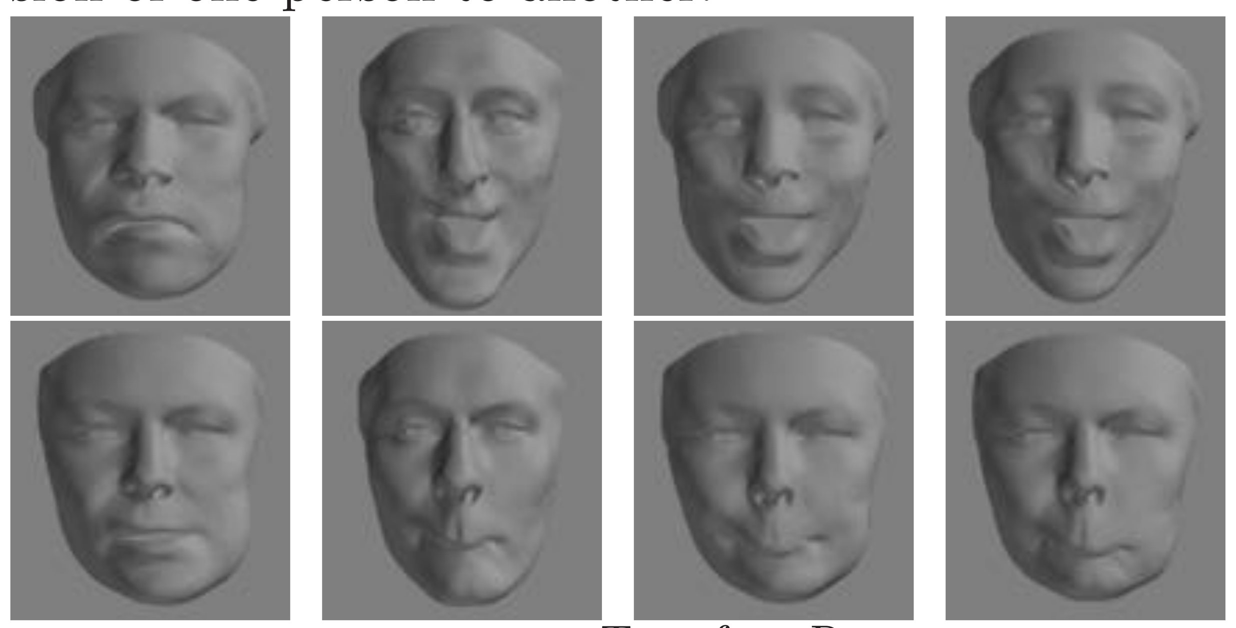
Method	Mean angular error against [5]
[1]	38.35°± 15.63°
Ours	**33.37°± 3.29°**

6. Expression and Identity

$$\boldsymbol{X} = \boldsymbol{B}_{(1)}(\boldsymbol{E} \odot \boldsymbol{C}) \quad (5)$$

Input Data: 3D synthetic data of faces with changes in identity and expressions

Result: Our method is able to disentangle identity and expression such that we can transfer the expression of one person to another.



Person | Expression | Transfer Result | Ground Truth

7. In-the-Wild Reconstruction

Input Data: "in-the-wild" images of faces

Result: Our method is able to reconstruct 3D shape even in challenging conditions such as noise and occlusions.

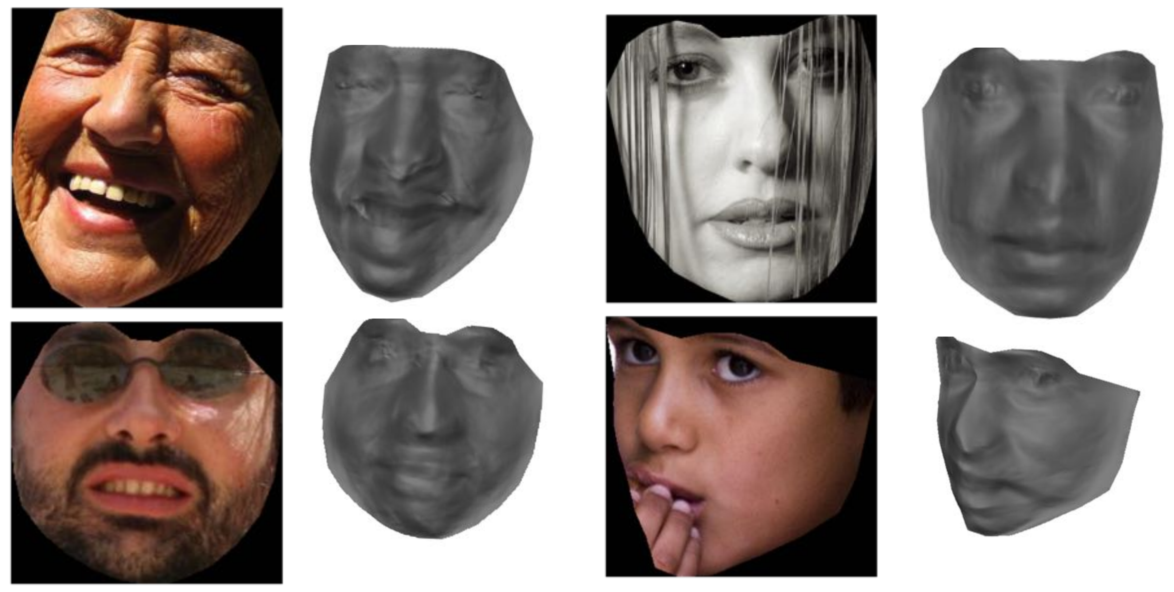


Image | Result | Image | Result

Input Data: "in-the-wild" images of ears

Result: Our method also reconstructs the 3D shape of ears.

Image | Result | Image | Result

8. Light, Expression, Identity

$$\boldsymbol{X} = \boldsymbol{B}_{(1)}(\boldsymbol{L} \odot \boldsymbol{E} \odot \boldsymbol{C}) \quad (6)$$

Input Data: Images of different people, different expressions and illumination changes

Result: Our method disentangles illumination, identity and expression. We can reconstruct their 3D shape and transfer the expression of one person to another.

Person | Expression | Transferred Expression | Ground Truth

References

[1] R. Basri, D. Jacobs, and I. Kemelmacher. Photometric stereo with general, unknown lighting. *International Journal of Computer Vision*, 2007.

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

[3] 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.

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