Compositional Hierarchical Tensor Factorization: Representing Hierarchical Intrinsic and Extrinsic Causal Factors

M. Alex O. Vasilescu[2,1]
maov@cs.ucla.edu

Eric Kim[1,2]
ekim@cs.ucla.edu

[1] Department of Computer Science
University of California, Los Angeles

[2] Tensor Vision Technologies
Los Angeles, CA, USA

ABSTRACT

Visual objects are composed of a recursive hierarchy of perceptual wholes and parts, whose properties, such as shape, reflectance, and color, constitute a hierarchy of intrinsic causal factors of object appearance. However, object appearance is the compositional consequence of both an object's intrinsic and extrinsic causal factors, where the extrinsic causal factors are related to illumination, and imaging conditions.Therefore, this paper proposes a unified tensor model of wholes and parts, and introduces a compositional hierarchical tensor factorization that disentangles the hierarchical causal structure of object image formation, and subsumes multilinear block tensor decomposition as a special case. The resulting object representation is an interpretable combinatorial choice of wholes' and parts' representations that renders object recognition robust to occlusion and reduces training data requirements. We demonstrate our approach in the context of face recognition by training on an extremely reduced dataset of synthetic images, and report encouraging face verification results on two datasets – the Freiburg dataset, and the Labeled Face in the Wild (LFW) dataset consisting of real-world images, thus, substantiating the suitability of our approach for data starved domains.

CCS CONCEPTS

• **Computing methodologies → Causal reasoning and diagnostics**; **Factorization methods**; **Factor analysis**; **Hierarchical representations**; **Learning latent representations**; **Object recognition**; **Image representations**; **Appearance and texture representations**; *Principal component analysis*; *Biometrics*; • **Mathematics of computing → Exploratory data analysis**; **Dimensionality reduction**; *Computations on matrices*; **Multivariate statistics**;

ACM Reference Format:
M.A.O. Vasilescu and E. Kim. 2019. Compositional Hierarchical Tensor Factorization: Representing Hierarchical Intrinsic and Extrinsic Causal Factors . In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'19): Tensor Methods for Emerging Data Science Challenges, August 04-08, 2019, Anchorage, AK.* ACM, New York, NY, USA, 9 pages.

KDD '19: TMEDSC, August 4-8, 2019, Anchorage, AK, USA

ACM ISBN 978-1-4503-6201-6/19/08.

1 INTRODUCTION

Statistical data analysis that disentangles the causal factors of data formation and computes a representation that facilitates the analysis, visualization, compression, approximation, and/or interpretation of the data is challenging and of paramount importance.

"Natural images are the composite consequence of multiple constituent factors related to scene structure, illumination conditions, and imaging conditions. Multilinear algebra, the algebra of higher-order tensors, offers a potent mathematical framework for analyzing the multifactor structure of image ensembles and for addressing the difficult problem of disentangling the constituent factors or modes." (Vasilescu and Terzopoulos, 2002 [69])

Scene structure is composed from a set of objects that appear to be formed from a recursive hierarchy of perceptual wholes and parts whose properties, such as shape, reflectance, and color, constitute a hierarchy of intrinsic causal factors of object appearance. Object appearance is the compositional consequence of both an object's intrinsic causal factors, and extrinsic causal factors with the latter related to illumination (i.e. the location and types of light sources), imaging (i.e. viewpoint, viewing direction, lens type and other camera characteristics). Intrinsic and extrinsic causal factors confound each other's contributions hindering recognition.

"Intrinsic properties are by virtue of the thing itself and nothing else" (David Lewis, 1983 [42]); whereas an extrinsic properties are not entirely about that thing, but as result of the way the thing interacts with the world. Unlike global intrinsic properties, local intrinsic properties are intrinsic to a part of the thing, and it may be said that a local intrinsic property is in an "intrinsic fashion", or "intrinsically" about the thing, rather than "is intrinsic" to the thing [21, 31]. David Lewis [42] provides a formal discussion of intrinsic and extrinsic concepts of causality and addresses a few related distinctions that an intuitive definition conflates, such as local versus global intrinsic properties, duplication preserving properties, and interior versus exterior properties. The meaning of intrinsic and extrinsic causation was extensively explored in philosophy, philosophy of mind, metaphysics and philosophy of physics [32, 42, 43, 46, 52].

Our goal is to explicitly represent local and global intrinsic causal factors as statistically invariant representations to all other causal factors of data formation.

Historically, statistical object recognition paradigms can be categorized based on how object structure is represented and recognized, ie. based on the appearance of an object's local features [22, 75], or based on the overall global object appearance [2, 4, 26, 45, 56, 62, 78]. Both approaches have strengths and shortcomings. Global features are sensitive to occlusions, while local features are sensitive to local deformations and noise. A hybrid approach that employs both global object features, and local features mitigates the shortcoming of both approaches [44, 47, 77].

Deep learning methods, which have become a highly successful approach for object recognition, compute feature hierarchies composed of low-level and mid-level features either in a supervised [41], unsupervised [27, 39, 54] or semi-supervised manner [48]. This has been achieved by composing modules of the same architectures, such as Restricted Boltzmann Machines [27], autoencoders [39], or various forms of encoder-decoder networks [8, 33, 54].

Cohen *et al.*'s [11, 12] theoretical results show that convolutional neural networks (CNNs) are theoretically equivalent in their representational power to hierarchical Tucker factorization [23, 24, 50, 53],[1] and shallow networks are equivalent to linear tensor factorizations, aka CANDECOMP/Parafac (CP) tensor factorization [7, 9, 25]. Vasilescu and Terzopoulos [63][69][70][72][67] demonstrated that tensor algebra is a suitable,[2] interpretable framework for mathematically representing and disentangling the causal structure of data formation in computer vision, computer graphics and machine learning [66, 72], Figure 1. Problem setup and implementation differences between CNNs and our tensor algebraic approach impact interpretability, data needs, memory/storage and computational complexity rendering CNN models difficult to deploy on mobile devices, or any devices with limited computational resources.

Inspired and inspirited by Cohen *et al.*'s [11, 12] theoretical results, and by the TensorFaces and Human Motion Signatures approach [63, 68–70], we propose a unified tensor model of wholes and parts based on a reconceptualization of the data tensor as a hierarchical data tensor, a mathematical representation of a tree datastructure. Defining a hierarchical data tensor enables a single elegant mathematical model that can be optimized in a principled manner instead of employing a myriad of individual part-based engineering solutions that independently represent each part and attempt to compute all possible dimensionality reduction permutations. Our factorization optimizes simultaneously across all the wholes and parts of the hierarchy, learns a convolutional feature hierarchy of low-level, mid-level and high-level features, and computes an interpretable compositional object representation of parts and wholes. Our resulting object representation is a combinatorial choice of part representations, that renders object recognition robust to occlusion while bypassing large training data requirements.

Our compositional tensor factorization and learnt feature hierarchy is also applicable to CNNs. Since CNNs learn millions of parameters that may lead to redundancy and poor generalization, our factorization and dimensionality reduction approach may be employed to reparameterize and reduce a tensor of CNN parameters, potentially resulting in better generalization [35, 38, 40, 49].

DeepFace, a CNN approach [10, 29, 59, 60, 76], celebrated closing the gap on human-level performance for face verification by testing on the Labeled Faces in the Wild (LFW) database [30] of 13,233 images from 5,749 people in the news, and training on a large dataset of 4,400,000 facial images from 4,030 people, the same order of magnitude as the number of people in the test data set [51]. This large training data resulted in an increase from 70% verification rates to 97.35%.

[1]While the Tucker factorization performs one matrix SVD to represent a causal factor subspace (one orthonormal mode matrix), a Hierarchical Tucker is a hierarchical computation of the Tucker factorization that employs a stack of SVDs to represent a causal factor subspace (one orthonormal mode matrix). For computational efficiency, the authors [23, 24] prescribe a stack of QR decompositions instead of a stack of SVDs.

[2]The suitability of the tensor framework was also demonstrated in the context of computer graphics by synthesizing new textures [71], performing expression retargeting/reanimation [74] and by synthesizing human motion [64][65][28].

The very resources that make deep learning a wildly successful approach today are also its shortcomings. In general, it is difficult and expensive to acquire large representative training data for object image analysis or recognition, and once acquired, there is a need for high performance computing, such as distributed GPU computing [1, 3, 5, 13].

While we have not closed the gap on human performance, the expressive power of our representation and our verification results are promising. We demonstrate and validate our novel compositional tensor representation in the context of the face verification, albeit it is intended for any data verification or classification scenario. We trained on less than one percent (1%) of the total images used by DeepFace. We trained on the images of only 100 people and tested on the images of the 5,749 people in the LFW database.

Contributions:

1. This paper (i) explicitly addresses the meaning of intrinsic versus extrinsic causality, and (ii) models cause-and-effect as multilinear tensor interaction between intrinsic and extrinsic hierarchical causal factors of data formation. The causal factor representations are interpretable, hierarchical, statistically invariant to all other causal factors and computed based on 2^{nd} order statistics, but may be extended to employ higher-order statistics, or kernel approaches.

2. In analogy to autoencoders which are inefficient and approximate neural network implementations of principal component analysis, a pattern analysis method based in linear algebra, CNNs are neural network implementations of tensor factorizations.

This paper contributes to the tensor algebraic paradigm: **(i)** we express our data tensor in terms of a unified tensor model of wholes and parts by defining a hierarchical data tensor; **(ii)** we introduce a compositional hierarchical tensor factorization that subsumes block-tensor decomposition as a special case [16, 19]; **(iii)** we validate our approach by employing our new compositional hierarchical tensor factorization in the context of face recognition, but it may be applied to any type of data.This approach is data agnostic.

2 RELEVANT TENSOR ALGEBRA

We will use standard textbook notation,denoting scalars by lower case italic letters ($a, b, ...$), vectors by bold lower case letters ($\mathbf{a}, \mathbf{b}, ...$), matrices by bold uppercase letters ($\mathbf{A}, \mathbf{B}, ...$), and higher-order tensors by bold uppercase calligraphic letters ($\mathcal{A}, \mathcal{B}, ...$). Index upper bounds are denoted by italic uppercase (*i.e.*, $1 \le i \le I$). The zero matrix is denoted by $\mathbf{0}$, and the identity matrix is denoted by $\mathbf{I}$.

Briefly, a *tensor*, or m-way array, is a generalization of a vector (first-order tensor) and a matrix (second-order tensor). Tensors are multilinear mappings over a set of vector spaces. The *order* of tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_M}$ is M. An element of $\mathcal{A}$ is denoted as $\mathcal{A}_{i_1 \ldots i_m \ldots i_M}$ or $a_{i_1 \ldots i_m \ldots i_M}$, where $1 \le i_m \le I_m$. In tensor terminology, column vectors are referred to as mode-1 vectors and row vectors as mode-2 vectors. The mode-m vectors of an M^{th}-order tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_M}$ are the I_m-dimensional vectors obtained from $\mathcal{A}$ by varying index i_m while keeping the other indices fixed. The mode-m vectors of a tensor are also known as *fibers*. The mode-m vectors are the column vectors of matrix $\mathbf{A}_{[m]}$ that results from *matrixizing* (a.k.a. *flattening*) the tensor $\mathcal{A}$ (Fig. 2).

Definition 1 (Mode-m Matrixizing). *The mode-m matrixizing of tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots I_M}$ is defined as the matrix $\mathbf{A}_{[m]} \in \mathbb{R}^{I_m \times (I_1 \ldots I_{m-1} I_{m+1} \ldots I_M)}$. As the parenthetical ordering indicates, the mode-m column vectors are arranged by sweeping all the other*

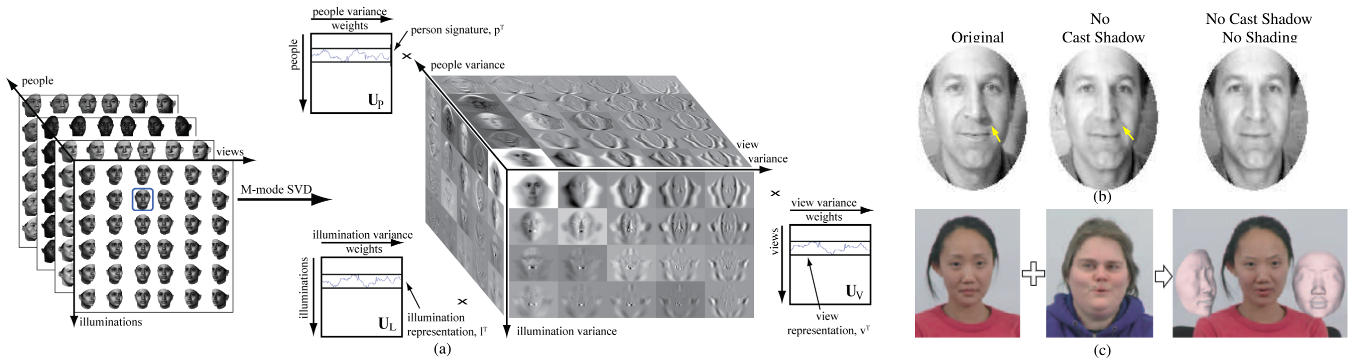

Figure 1: (a) TensorFaces representation: This image illustrates a multilinear tensor factorization of a 4^{th}-order training data tensor, $\mathcal{D} \in \mathbb{R}^{I_x \times I_L \times I_V \times I_P}$ where I_x, I_L, I_V, I_P are the number of pixels, illuminations, views and people, respectively. A vectorized image, d, is represented by a set of coefficient vectors, one for the illumination, viewpoint, and person, and expressed mathematically as $\mathrm{d} = \mathcal{T} \times_L \mathrm{l}^T \times_V \mathrm{v}^T \times_P \mathrm{p}^T$, where the TensorFaces basis, $\mathcal{T}$, governs the interaction between different causal factors of data formation. (Data should be centered, but for this display it was added back.) (b) These set of images demonstrate the models ability to disentangle the causal factors. Illumination effects, such as highlights, cast shadows and shading [70] are predectively and progressively reduced. (c) Vlasic etal [74] generate new images by performing multilinear expression re-targeting.

mode indices through their ranges, with smaller mode indexes varying more rapidly than larger ones; thus,

$$\left[\boldsymbol{A}_{[m]}\right]_{jk} = a_{i_1 \ldots i_m \ldots i_M}, \quad \textit{where} \tag{1}$$

$$j = i_m \quad \textit{and} \quad k = 1 + \sum_{\substack{n=1 \\ n \neq m}}^{M} (i_n - 1) \prod_{\substack{l=1 \\ l \neq m}}^{n-1} I_l.$$

A generalization of the product of two matrices is the product of a tensor and a matrix [17].

Definition 2 (Mode-m Product, $\times_M$). *The mode-m product of a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_m \times \cdots \times I_M}$ and a matrix $\boldsymbol{B} \in \mathbb{R}^{J_m \times I_m}$, denoted by $\mathcal{A} \times_m \boldsymbol{B}$, is a tensor of dimensionality $\mathbb{R}^{I_1 \times \cdots \times I_{m-1} \times J_m \times I_{m+1} \times \cdots \times I_M}$ whose entries are*

$$[\mathcal{A} \times_m \boldsymbol{B}]_{i_1 \ldots i_{m-1} j_m i_{m+1} \ldots i_M} = \sum_{i_m} a_{i_1 \ldots i_{m-1} i_m i_{m+1} \ldots i_M} b_{j_m i_m}.$$

The mode-m product can be expressed in tensor notation, as $\mathcal{C} = \mathcal{A} \times_m \mathbf{B}$, or in terms of matrixized tensors, as $\mathbf{C}_{[m]} = \mathbf{B}\mathbf{A}_{[m]}$. The M-mode SVD (aka. the Tucker decomposition) is a "generalization" of the conventional matrix (i.e., 2-mode) SVD which may be written in tensor notation as

$$\mathbf{D} = \mathbf{U}_1 \mathbf{S} \mathbf{U}_2^T \quad \Leftrightarrow \quad \mathbf{D} = \mathbf{S} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2. \tag{2}$$

The M-mode SVD orthogonalizes the M spaces and decomposes the tensor as the *mode-m product*, denoted $\times_m$, of M-orthonormal spaces, as follows:

$$\mathcal{D} = \mathcal{Z} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \cdots \times_m \mathbf{U}_m \cdots \times_M \mathbf{U}_M. \tag{3}$$

3 GLOBAL TENSOR FACTORIZATION

There are two classes of data tensor modeling techniques [37, 58] that stem from: the rank-K decomposition (CANDECOMP/Parafac decomposition) [7, 9, 25] and the multilinear rank-$(R_1, R_2, \ldots, R_M)$, such as Tucker decomposition [61][17][20], such as Multilinear-PCA, multilinear (tensor) ICA, plus various kernel variations that are doubly nonlinear [66].

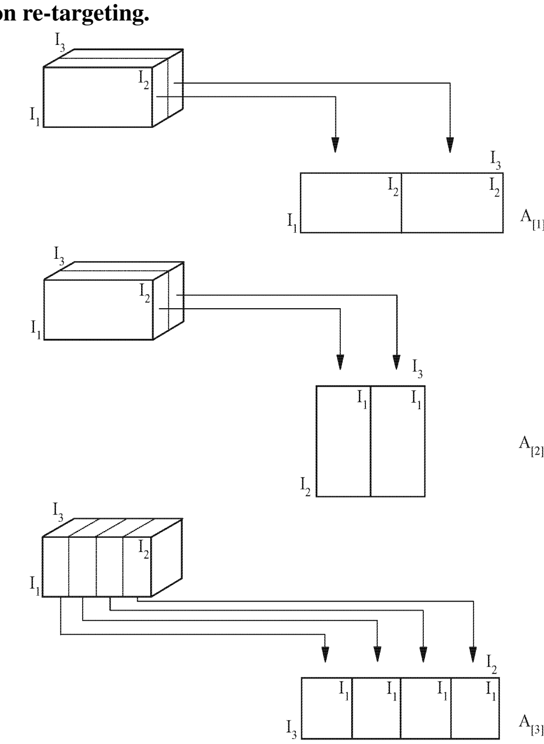

Figure 2: Matrixizing a (3rd-order) tensor. The tensor can be matrixized in 3 ways to obtain matrices comprising its 1-mode, 2-mode, and 3-mode vectors. Note that this matrixizing is not cyclical, unlike the one defined in [17].

Algorithm 1 M-mode SVD algorithm.

Input the data tensor $\mathcal{D} \in \mathbb{R}^{I_1 \times \cdots \times I_M}$.

(1) For $m := 1, \ldots, M$,
Let $\mathbf{U}_m$ be the left orthonormal matrix of the SVD of $\mathbf{D}_{[m]}$, the mode-m matrixized $\mathcal{D}$.

(2) Set $\mathcal{Z} := \mathcal{D} \times_1 \mathbf{U}_1^T \times_2 \mathbf{U}_2^T \cdots \times_m \mathbf{U}_m^T \cdots \times_M \mathbf{U}_M^T$.

Output mode matrices $\mathbf{U}_1, \ldots, \mathbf{U}_M$ and the core tensor $\mathcal{Z}$.

3.1 Representation: Multilinear Tensor Factorization

Within the tensor mathematical framework, an ensemble of observations is organized in a higher order data tensor, $\mathcal{D}$. Given a data tensor $\mathcal{D}$ of labeled, vectorized training facial images $\mathbf{d}_{pvle}$, where the subscripts denote the causal factors of facial image formation, the person p, view v, illumination l, and expression e labels, the M-mode SVD [70][65][17][18] or its kernel variant [66] may be employed to multilinearly decompose the data tensor,

$$\mathcal{D} = \mathcal{T} \times_{\mathrm{P}} \mathbf{U}_{\mathrm{P}} \times_{\mathrm{V}} \mathbf{U}_{\mathrm{V}} \times_{\mathrm{L}} \mathbf{U}_{\mathrm{L}} \times_{\mathrm{E}} \mathbf{U}_{\mathrm{E}}, \tag{4}$$

and compute the mode matrices $\mathbf{U}_{\mathrm{P}}$, $\mathbf{U}_{\mathrm{V}}$, $\mathbf{U}_{\mathrm{L}}$, and $\mathbf{U}_{\mathrm{E}}$ that span the causal factor representation. The extended core computed by, $\mathcal{T} = \mathcal{D} \times_{\mathrm{P}} \mathbf{U}_{\mathrm{P}}^{\mathrm{T}} \times_{\mathrm{V}} \mathbf{U}_{\mathrm{V}}^{\mathrm{T}} \times_{\mathrm{L}} \mathbf{U}_{\mathrm{L}}^{\mathrm{T}} \times_{\mathrm{E}} \mathbf{U}_{\mathrm{E}}^{\mathrm{T}}$, governs the interaction between the causal factors (Figure 1). This approach makes the assumption that the causal factor representations are well modeled by a Gaussian distribution. An image $\mathbf{d}_{pvle}$ is represented by a person, view, illumination and expression coefficient vectors as,

$$\mathbf{d}_{pvle} = \mathcal{T} \times_{\mathrm{P}} \mathbf{p}_p^{\mathrm{T}} \times_{\mathrm{V}} \mathbf{v}_v^{\mathrm{T}} \times_{\mathrm{L}} \mathbf{l}_l^{\mathrm{T}} \times_{\mathrm{E}} \mathbf{e}_e^{\mathrm{T}}. \tag{5}$$

An important advantage of employing vectorized observations in a multilinear tensor framework is that all images of a person, regardless of viewpoint, illumination and expression are mapped to the same person coefficient vector, thereby achieving zero intra-class scatter. Thus, multilinear analysis creates well separated people classes by maximizing the ratio of inter-class scatter to intra-class scatter. Alternatively, one can employ the Multilinear (Tensor) Independent Component Analysis (MICA) algorithm [72], M-mode ICA (as opposed to the tensorized computation of the conventional, linear ICA[15]), which takes advantage of higher-order statistics [14] to compute the mode matrices that span the causal factor representation, and the MICA basis tensor that governs their interaction.

3.2 Recognition: Mutilinear Projection

While TensorFaces (MPCA) [69, 70] is a handy moniker for an approach that learns from an image ensemble the interaction and representation of various causal factors that determine observed data, with Multilinear (Tensor) ICA [72] as a more sophisticated approach, none of the interaction models prescribe a solution for how one might determine the causal factors of a single unlabeled test image that is not part of the training set. Multilinear projection [67, 73] simultaneously projects one or more unlabeled test images that are not part of the training data set into multiple constituent causal factor spaces associated with data formation, in order to infer the mode labels:

$$\mathcal{R} = \mathcal{T}^{+_{\mathrm{x}}} \times_{\mathrm{x}}^{\mathrm{T}} \mathbf{d}_{\mathrm{new}} \approx \mathbf{r}_{\mathrm{P}} \circ \mathbf{r}_{\mathrm{V}} \circ \mathbf{r}_{\mathrm{L}} \circ \mathbf{r}_{\mathrm{E}}. \tag{6}$$

The multilinear projection of a facial image computes the illumination, view and person representation by decomposing the expected rank-1 structure of the confounded response tensor $\mathcal{R}$, and taking advantage of the unit vector constraints associated with $\mathbf{r}_{\mathrm{L}}$, $\mathbf{r}_{\mathrm{V}}$, $\mathbf{r}_{\mathrm{E}}$, and $\mathbf{r}_{\mathrm{P}}$. via the CP-decomposition, *i.e.*, the rank-1 tensor decomposition.

4 COMPOSITIONAL HIERARCHICAL TENSOR FACTORIZATIONS

In prior tensor based research, an imaged object was represented in terms of global causal factor representations that are not robust to occlusions. This section introduces a compositional hierarchical tensor factorization that derives its name from its ability to represent a hierarchy of intrinsic and extrinsic causal factors of data formation.A hierarchical representation may be effectively employed to recognize occluded objects, including self-occlusion that occurs during out-of-plane rotation relative to the camera viewpoint. The efficacy of our approach is demonstrated by our LFW and Freiburg verification experiments that compare global and hierarchical representations in Section 5.

Within the tensor mathematical framework, an ensemble of training observations is organized in a higher order data tensor, $\mathcal{D}$. A data tensor $\mathcal{D} \in \mathbb{R}^{I_0 \times I_1 \times I_c \cdots \times I_C}$ contains a collection of vectorized observation, $\mathbf{d}_{i_1 \ldots i_c \ldots i_C} \in \mathbb{R}^{I_0}$ where each subscript, i_c, denotes one of the C causal factors that have created the observation and have resulted in I_0 measurements, *i.e.*, a total of I_0 pixels per image. In this paper, we will report results based on a data tensor $\mathcal{D}$ of labeled, vectorized training facial images $\mathbf{d}_{pvle}$, where the subscripts denote the causal factors of facial image formation, the person p, view v, illumination l, and expression e labels.

4.1 Hierarchical Data Tensor

We identify a general base case object and two special cases whose intrinsic and extrinsic causal factors we would like to represent. An object may be composed of (i) two partially overlapping children-parts and parent-whole that has data not contained in any of the children-parts, (ii) a set of non-overlapping parts, or (iii) a set of fully overlapping parts, which resembles to the rank-(L, M, N) or a rank-$(L, M, \cdot)$ block tensor decomposition [16], but which is too restrictive for our purpose.[3] Figure 3 depicts the general base case and the two special base cases. In real scenarios, parent-wholes have children-parts that are recursively composed of children themselves, Fig. 4.

The data wholes, and parts are extracted by employing a filter bank, $\{\mathbf{H}_{\mathrm{s}} \| \mathbf{H}_{\mathrm{s}} \in \mathbb{C}^{J_0 \times I_0}, 1 \le s \le S\}$ where each $2D$ convolutional filter is implemented as a doubly (triply) block circulant matrix,$\mathbf{H}_{\mathrm{s}}$, where $\sum_{s=1}^{S} \mathbf{H}_{\mathrm{s}} = \mathbf{I}$, and s refers to the data segment. Convolution is a matrix-vector multiplication, between a circulant matrix and a vectorized observation, $\mathbf{d}_{\mathrm{s}} = \mathbf{H}_{\mathrm{s}}\mathbf{d}$ which in tensor notation is written as $\mathbf{d}_{\mathrm{s}} = \mathbf{d} \times_0 \mathbf{H}_{\mathrm{s}}$ where mode 0 is the measurement mode. The segment data tensor, $\mathcal{D}_S = \mathcal{D} \times_0 \mathbf{H}_{\mathrm{s}}$, is the result of multiplying (convolving) every observation, $\mathbf{d}$, with the block circulant matrix (filter), $\mathbf{H}_{\mathrm{s}}$. A filter $\mathbf{H}_{\mathrm{s}}$ may be any of any type, or have any spatial scope. When a filter matrix is a block identity matrix, $\mathbf{H}_{\mathrm{s}} = \mathbf{I}_{\mathrm{s}}$, the filter matrix multiplication with a vectorized observation has the effect of segmenting a portion of the data without any blurring, subsampling or upsampling.Measurements associated with perceptual parts may not be tightly packed into a block a priori, as in the case of vectorized images, but chunking may be achieved by a trivial permutation.

The data tensor, $\mathcal{D}$, is expressed in terms of its recursive hierarchy of wholes and parts by defining and employing a *hierarchical data tensor*, $\mathcal{D}_{\mathcal{H}}$, that contains along its super-diagonal the collection of wholes and parts, $\mathcal{D}_{\mathrm{s}}$, Fig. 3(a)

$$\begin{aligned} \mathcal{D} &= \sum_{s=1}^{S} \mathcal{D} \times_0 \mathbf{H}_{\mathrm{s}} && (7)\\ &= \mathcal{D}_1 \cdots + \mathcal{D}_{\mathrm{s}} \cdots + \mathcal{D}_{\mathrm{S}} && (8)\\ &= \mathcal{D}_{\mathcal{H}} \times_0 \mathbf{I}_{0\mathrm{x}} \times_1 \mathbf{I}_{1\mathrm{x}} \cdots \times_c \mathbf{I}_{c\mathrm{x}} \cdots \times_C \mathbf{I}_{C\mathrm{x}}, && (9) \end{aligned}$$

[3] The block tensor decomposition [16] computes the best fitting R fully overlaping tensor blocks that are all multilinearly decomposable into the same multilinear rank-(L,M,N),which is analogous to finding the best fitting R rank-1 terms computed by the CP-algorithm.

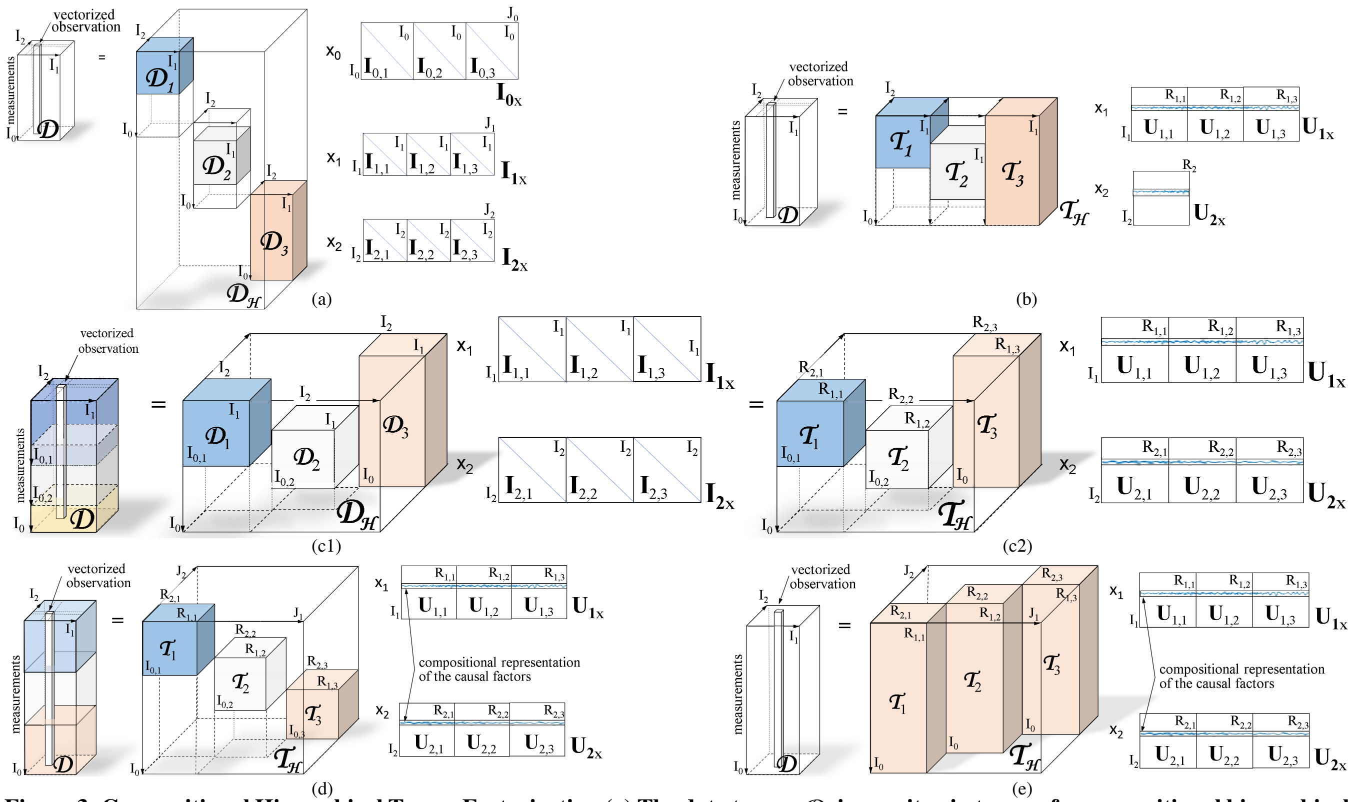

Figure 3: Compositional Hierarchical Tensor Factorization.(a) The data tensor, $\mathcal{D}$, is rewriten in terms of a compositional hierarchical data tensor, $\mathcal{D}_\mathcal{H}$. A *fully compositional hierarchical data tensor* is a data tensor in which every mode is written in a compositional form. The compositional hierarchical data tensor, $\mathcal{D}_\mathcal{H}$, contains data tensor segments along its super-diagonal. The data tensor segment, $\mathcal{D}_s$, may contain zeros and represent parts, or may be full and correspond to a filtered version of a parent-whole. (b) Only one of the causal factor has a compositional hierarchical representation, a *partially compositional hierarchical* representation example.(c) In practice, the measurement mode will not be written in a compositional form, ie. $\mathcal{D}_\mathcal{H}$ would have already been multiplied with $\mathbf{I}_{0x}$as in Fig. 3(c1). (c1) Rewriting the data tensor $\mathcal{D}$ as a hierarchical data tensor, $\mathcal{D}_\mathcal{H}$. (c2) Performing a compositional hierarchical tensor factorization results in a part-based causal factor representation, $\mathcal{D} = \mathcal{T}_\mathcal{H} \times_1 \mathbf{U}_{1x} \times_2 \mathbf{U}_{2x}$, where the extended core is $\mathcal{T}_\mathcal{H} = \mathcal{Z}_\mathcal{H} \times_0 \mathbf{U}_{0x}$.(d) Non-overlapping parts. The part representations' computations are independent of one another. (e) Completely overlapping parts.

where $\mathbf{I}_{cx} = [\mathbf{I}_{c,1} \ldots \mathbf{I}_{c,s} \ldots \mathbf{I}_{c,S}] \in \mathbb{R}^{I_c \times SI_c}$ is a concatenation of S identity matrices, one for each data segment. The three different ways of rewritting $\mathcal{D}$ in terms of a hierarchy of wholes and parts, eq. 7-9, results in three equivalent compositional hierarchical tensor factorizations [4]

$$\mathcal{D} = \sum_{s=1}^{S} \underbrace{(\mathcal{Z} \times_0 \mathbf{U}_0 \times_1 \mathbf{U}_1 \cdots \times_c \mathbf{U}_c \cdots \times_C \mathbf{U}_C)}_{\mathcal{D}} \times_0 \mathbf{H}_s \tag{10}$$

$$= \sum_{s=1}^{S} \underbrace{(\mathcal{Z}_s \times_0 \mathbf{U}_{0,s} \times_1 \mathbf{U}_{1,s} \cdots \times_c \mathbf{U}_{c,s} \cdots \times_C \mathbf{U}_{C,s})}_{\mathcal{D}_s} \tag{11}$$

$$= \underbrace{(\mathcal{Z}_\mathcal{H} \times_0 \mathbf{U}_{0\mathcal{H}} \times_1 \mathbf{U}_{1\mathcal{H}} \cdots \times_c \mathbf{U}_{c\mathcal{H}} \cdots \times_C \mathbf{U}_{C\mathcal{H}})}_{\mathcal{D}_\mathcal{H}}$$

$$\times_0 \mathbf{I}_{0x} \times_1 \mathbf{I}_{1x} \cdots \times_c \mathbf{I}_{cx} \cdots \times_C \mathbf{I}_{Cx} \tag{12}$$

The expression of $\mathcal{D}$ in terms of a hierarchical data tensor is a mathematical conceptual device, that enables a unified mathematical model of wholes and parts that can be expressed completely as a mode-m product (matrix-vector multiplication) and whose factorization can be optimized in a principled manner. Dimensionality reduction of the compositional representation is performed by optimizing

$$e = \|\mathcal{D} - (\bar{\mathcal{Z}}_\mathcal{H} \times_0 \bar{\mathbf{U}}_{0\mathcal{H}} \ldots \times_c \bar{\mathbf{U}}_{c\mathcal{H}} \ldots \times_C \bar{\mathbf{U}}_{C\mathcal{H}})$$

$$\times_0 \mathbf{I}_{0x} \ldots \times_c \mathbf{I}_{cx} \ldots \times_C \mathbf{I}_{Cx}\|^2 + \sum_{c=0}^{C} \lambda_c \|\bar{\mathbf{U}}_{c\mathcal{H}}^{\mathrm{T}} \bar{\mathbf{U}}_{c\mathcal{H}} - \mathbf{I}\|^2 \tag{13}$$

where $\bar{\mathbf{U}}_{c\mathcal{H}}$ is the composite representation of the c^{th} mode, and $\mathcal{Z}_\mathcal{H}$ governs the interaction between causal factors.[5] Our optimization may be initialized, only, by setting $\mathcal{Z}_\mathcal{H}$ and $\mathbf{U}_{c\mathcal{H}}$ to the M-mode SVD of $\mathcal{D}_\mathcal{H}$, [6] and performing dimensionality reduction through truncation, where $\bar{\mathbf{U}}_{c\mathcal{H}} \in \mathbb{R}^{SI_c \times \bar{J}_c}$, $\bar{\mathcal{Z}}_\mathcal{H} \in \mathbb{R}^{\bar{J}_0 \cdots \times \bar{J}_c \cdots \times \bar{J}_C}$ and $\bar{J}_c \leq SI_c$.

[4]Equivalent representations can be transformed into one another by post-multiplying mode matrices with permutations or more generally nonsingular matrices, $\mathbf{G}_c$, $\mathcal{D} = (\mathcal{Z}_\mathcal{H} \times_0 \mathbf{G}_0^{-1} \cdots \times_c \mathbf{G}_c^{-1} \cdots \times_C \mathbf{G}_C^{-1}) \times_1 \mathbf{I}_{1x}\mathbf{U}_{1\mathcal{H}}\mathbf{G}_1 \cdots \times_c \mathbf{I}_{cx}\mathbf{U}_{c\mathcal{H}}\mathbf{G}_c \cdots \times_C \mathbf{I}_{Cx}\mathbf{U}_{C\mathcal{H}}\mathbf{G}_C$.

[5]In the face recogniton application we discuss later, $c = 0$ refers to the measurement mode, *i.e.*, the pixel values in an image, and the range of values $1 \leq c \leq C$ refers to the C causal factors.

[6]For computational efficiency, we may perform M-mode SVD on each data tensor segment $\mathcal{D}_s$ and concatenate terms along the diagonal of $\mathcal{Z}_\mathcal{H}$ and $\mathbf{U}_{c\mathcal{H}}$. However, the most computational efficient initialization, first, computes the M-mode SVD of $\mathcal{D}$, multiplies (*i.e.*,,convolves) the core tensor, $\mathcal{Z}$ with $\mathbf{H}_s$, followed by a concatenation of terms along the diagonal of $\mathcal{Z}_\mathcal{H}$ and duplication of $\mathbf{U}_c$ along the diagonal of $\mathbf{U}_{c\mathcal{H}}$. The last initialization approach makes segment specific dimensionality reduction problematic, since part-based standard deviation, $\sigma_{i_c i_c,s}$, is not computed.

4.2 Compositional Hierarchical Factorization Derivation

For notational simplicity, we re-write the loss function as,

$$\begin{aligned} e := &\|\mathcal{D} - \tilde{\mathcal{Z}}_{\mathcal{H}} \times_0 \tilde{\mathbf{U}}_{0x} \ldots \times_c \tilde{\mathbf{U}}_{cx} \ldots \times_C \tilde{\mathbf{U}}_{Cx}\|^2 \\ &+ \sum_{c=0}^{C}\sum_{s=1}^{S} \lambda_{c,s}\|\tilde{\mathbf{U}}_{c,s}^{T}\tilde{\mathbf{U}}_{c,s} - \mathbf{I}\| \end{aligned} \qquad (14)$$

where $\tilde{\mathbf{U}}_{cx} = \mathbf{I}_{cx}\bar{\mathbf{U}}_{c\mathcal{H}}\tilde{\mathbf{G}}_{c} = [\tilde{\mathbf{U}}_{c,1}|\ldots|\tilde{\mathbf{U}}_{c,s}|\ldots|\tilde{\mathbf{U}}_{c,S}]$ and $\tilde{\mathbf{G}}_{c} \in \mathbb{R}^{\bar{J}_c \times SI_c}$ is permutation matrix that groups the columns of $\mathbf{U}_{c\mathcal{H}}$ based on the segment, s, to which they belong, and the inverse permutation matrices have been multiplied[4] into $\tilde{\mathcal{Z}}_{\mathcal{H}}$ resulting into a core that has also been grouped based on segments and sorted based on variance. The data tensor, $\mathcal{D}$, may be expressed in matrix form as in eq. 16 and reduces to the more efficiently block structure as in eq. 17

$$\mathcal{D} = \mathcal{Z}_{\mathcal{H}} \times_0 \mathbf{U}_{0x} \times_1 \mathbf{U}_{1x} \cdots \times_c \mathbf{U}_{cx} \cdots \times_C \mathbf{U}_{Cx} \qquad (15)$$

$$\begin{aligned} \mathbf{D}_{[c]} &= \mathbf{U}_{cx}\mathbf{Z}_{\mathcal{H}[c]}\left(\mathbf{U}_{Cx} \otimes \cdots \otimes \mathbf{U}_{(c+1)x} \otimes \mathbf{U}_{(c\text{-}1)x} \otimes \cdots \otimes \mathbf{U}_{0x}\right)^{T} && (16) \\ &= [\,\mathbf{U}_{c,1} \ldots \mathbf{U}_{c,s} \ldots \mathbf{U}_{c,S}\,] && (17) \\ & \quad \begin{bmatrix} \mathbf{Z}_{0[c]} & \mathbf{0} & & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} & & \vdots \\ \vdots & \mathbf{0} & \mathbf{Z}_{s[c]} & \mathbf{0} & \\ & & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & & \mathbf{0} & \mathbf{Z}_{S[c]} \end{bmatrix} \underbrace{\begin{bmatrix} \left(\mathbf{U}_{C,1} \ldots \otimes \mathbf{U}_{(c+1),1} \otimes \mathbf{U}_{(c\text{-}1),1} \ldots \otimes \mathbf{U}_{0,1}\right)^{T} \\ \vdots \\ \left(\mathbf{U}_{C,s} \ldots \otimes \mathbf{U}_{(c+1),s} \otimes \mathbf{U}_{(c\text{-}1),s} \ldots \otimes \mathbf{U}_{0,s}\right)^{T} \\ \vdots \\ \left(\mathbf{U}_{C,S} \ldots \otimes \mathbf{U}_{(c+1),S} \otimes \mathbf{U}_{(c\text{-}1),S} \ldots \otimes \mathbf{U}_{0,S}\right)^{T} \end{bmatrix}}_{\left(\mathbf{U}_{Cx} \cdots \odot \mathbf{U}_{(c+1)x} \odot \mathbf{U}_{(c\text{-}1)x} \cdots \odot \mathbf{U}_{0x}\right)^{T}} \\ &= \mathbf{U}_{cx}\mathbf{W}_{c}^{T}, && (18) \end{aligned}$$

where $\otimes$ is the Kronecker product,[7] and $\odot$ is the block-matrix Kahtri-Rao product.[8] The matricized block diagonal form of $\mathcal{Z}_{\mathcal{H}}$ in eq. 17 becomes evident when employing our modified data centric matrixizing operator based on the defintion 1, where the initial mode is the measurement mode.

The compositional hierarchical tensor factorization algorithm computes the mode matrix, $\mathbf{U}_{cx}$, by computing the minimum of $e = \|\mathcal{D} - \tilde{\mathcal{Z}}_{\mathcal{H}} \times_0 \tilde{\mathbf{U}}_{0x} \cdots \times_C \tilde{\mathbf{U}}_{Cx}\|^2$ by cycling through the modes, solving for $\tilde{\mathbf{U}}_{cx}$ in the equation $\partial e/\partial \mathbf{U}_{cx} = 0$ while holding the core tensor $\mathcal{Z}_{\mathcal{H}}$ and all the other mode matrices constant, and repeating until convergence. Note that

$$\frac{\partial e}{\partial \mathbf{U}_{cx}} = \frac{\partial}{\partial \mathbf{U}_{cx}}\|\mathbf{D}_{[c]} - \mathbf{U}_{cx}\mathbf{W}_{c}^{T}\|^2 = -\mathbf{D}_{[c]}\mathbf{W}_{c} + \mathbf{U}_{cx}\mathbf{W}_{c}^{T}\mathbf{W}_{c}. \qquad (19)$$

Thus, $\partial e/\partial \mathbf{U}_{cx} = 0$ implies that

$$\begin{aligned} \mathbf{U}_{cx} &= \mathbf{D}_{[c]}\mathbf{W}_{c}\left(\mathbf{W}_{c}^{T}\mathbf{W}_{c}\right)^{-1} = \mathbf{D}_{[c]}\mathbf{W}_{c}^{T+} && (20) \\ &= \mathbf{D}_{[c]}\left(\mathbf{Z}_{\mathcal{H}[c]}\left(\mathbf{U}_{Cx} \otimes \ldots \mathbf{U}_{(c+1)x} \otimes \mathbf{U}_{(c\text{-}1)x} \otimes \ldots \mathbf{U}_{0x}\right)^{T}\right)^{+} && (21) \\ &= \mathbf{D}_{[c]}\left(\mathbf{U}_{Cx} \odot \ldots \mathbf{U}_{(c+1)x} \odot \mathbf{U}_{(c\text{-}1)x} \odot \ldots \mathbf{U}_{0x}\right)^{T+} \begin{bmatrix} \mathbf{Z}_{0[c]}^{+} & \mathbf{0} & & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} & & \vdots \\ \vdots & \mathbf{0} & \mathbf{Z}_{s[c]}^{+} & \mathbf{0} & \\ & & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & & \mathbf{0} & \mathbf{Z}_{S[c]}^{+} \end{bmatrix} && (22) \end{aligned}$$

whose $\mathbf{U}_{c,s}$ sub-matrices are then subject to orthonormality constraints.

Solving for the optimal core tensor, $\mathcal{Z}_{\mathcal{H}}$, the data tensor, $\mathcal{D}$, approximation is expressed in vector form as,

$$e = \|\mathrm{vec}(\mathcal{D}) - (\tilde{\mathbf{U}}_{Cx} \otimes \cdots \otimes \tilde{\mathbf{U}}_{cx} \otimes \cdots \otimes \tilde{\mathbf{U}}_{0x})\mathrm{vec}(\tilde{\mathcal{Z}}_{\mathcal{H}})\|. \qquad (23)$$

Solve for the non-zero(nz) terms of $\mathcal{Z}_{\mathcal{H}}$ in the equation $\partial e/\partial(\mathcal{Z}_{\mathcal{H}}) = 0$, by removing the corresponding zero columns of the first matrix on right side of the equation below, performing the pseudo-inverse, and setting

$$\mathrm{vec}(\mathcal{Z}_{\mathcal{H}})_{nz} = (\mathbf{U}_{Cx} \otimes \cdots \otimes \mathbf{U}_{cx} \otimes \cdots \otimes \mathbf{U}_{0x})_{nz}^{+}\mathrm{vec}(\mathcal{D}). \qquad (24)$$

Repeat all steps until convergence.This optimization is the basis of the Compositional Hierarchical Tensor Factorization, Algorithm 2.

Completely Overlapping parts: When the data tensor is a collection of overlapping parts that have the same multilinear-rank reduction, Fig. 3e, the extended-core data tensor, $\mathcal{T}_{\mathcal{H}}$, computation in matrix form reduces to eq. 27

$$\begin{aligned} \mathcal{D} &= \mathcal{T}_{\mathcal{H}} \times_1 \mathbf{U}_{1x} \cdots \times_c \mathbf{U}_{cx} \cdots \times_C \mathbf{U}_{Cx} && (25) \\ \mathbf{D}_{[0]} &= \mathbf{T}_{\mathcal{H}[0]}(\mathbf{U}_{Cx} \cdots \otimes \mathbf{U}_{cx} \cdots \otimes \mathbf{U}_{1x})^{T} && (26) \end{aligned}$$

$$\begin{bmatrix} \mathbf{T}_{1[0]} \\ \vdots \\ \mathbf{T}_{s[0]} \\ \vdots \\ \mathbf{T}_{S[0]} \end{bmatrix}^{T} = \mathbf{D}_{[0]} \underbrace{\begin{bmatrix} (\mathbf{U}_{C,1} \cdots \otimes \mathbf{U}_{c,1} \cdots \otimes \mathbf{U}_{1,1})^{T} \\ \vdots \\ (\mathbf{U}_{C,s} \cdots \otimes \mathbf{U}_{c,s} \cdots \otimes \mathbf{U}_{1,s})^{T} \\ \vdots \\ (\mathbf{U}_{C,S} \cdots \otimes \mathbf{U}_{c,S} \cdots \otimes \mathbf{U}_{1,S})^{T} \end{bmatrix}^{+}}_{\left(\mathbf{U}_{Cx} \cdots \odot \mathbf{U}_{cx} \cdots \odot \mathbf{U}_{1x}\right)^{T+}}. \qquad (27)$$

Independent Parts: When the data tensor is a collection of observations made up of non-overlapping parts, Fig. 3d, the data tensor decomposition reduces to the concatenation of a M-mode SVD of individual parts,

$$\begin{bmatrix} \mathbf{D}_{1[c]} \\ \vdots \\ \mathbf{D}_{s[c]} \\ \vdots \\ \mathbf{D}_{S[c]} \end{bmatrix} = \begin{bmatrix} \mathbf{U}_{c,1}\mathbf{Z}_{0[c]}(\mathbf{U}_{C,1} \cdots \otimes \mathbf{U}_{(c+1),1} \otimes \mathbf{U}_{(c\text{-}1),1} \cdots \otimes \mathbf{U}_{0,1})^{T} \\ \vdots \\ \mathbf{U}_{c,s}\mathbf{Z}_{s[c]}(\mathbf{U}_{C,s} \cdots \otimes \mathbf{U}_{(c+1),s} \otimes \mathbf{U}_{(c\text{-}1),s} \cdots \otimes \mathbf{U}_{0,s})^{T} \\ \vdots \\ \mathbf{U}_{c,S}\mathbf{Z}_{S[c]}(\mathbf{U}_{C,S} \cdots \otimes \mathbf{U}_{(c+1),S} \otimes \mathbf{U}_{(c\text{-}1),S} \cdots \otimes \mathbf{U}_{0,S})^{T} \end{bmatrix} \qquad (28)$$

Note, that every $\mathbf{D}_{s[c]}$ row does not contain any terms from any other segment-part except segment s. Thus, every $\mathbf{U}_{c,s}$ and $\mathcal{Z}_{s}$ are computed by performing multilinear subspace learning, Algorithm 1, on the $\mathcal{D}_{s}$ and the results are appropriately concatenated in $\tilde{\mathbf{U}}_{cx}$ and $\tilde{\mathcal{Z}}_{\mathcal{H}}$.

[7] The Kronecker product of $\mathbf{U} \in \mathbb{R}^{IJ}$ and $\mathbf{V} \in \mathbb{R}^{KL}$ is the $IK \times JL$ matrix defined as $[\mathbf{U} \otimes \mathbf{V}]_{ik,jl} = u_{ij}v_{kl}$.

[8] The Khatri-Rao product of $[\mathbf{U}_1 \ldots \mathbf{U}_n \ldots \mathbf{U}_N] \odot [\mathbf{V}_1 \ldots \mathbf{V}_n \ldots \mathbf{V}_N]$ with $\mathbf{U}_l \in \mathbb{R}^{I \times N_l}$ and $\mathbf{V}_l \in \mathbb{R}^{K \times N_l}$ is a *block-matrix Kronecker product*; therefore, it can be expressed as $\mathbf{U} \odot \mathbf{V} = [(\mathbf{U}_1 \otimes \mathbf{V}_1) \ldots (\mathbf{U}_{(l)} \otimes \mathbf{V}_{(l)}) \ldots (\mathbf{U}_{(L)} \otimes \mathbf{V}_{(L)})]$ [55].

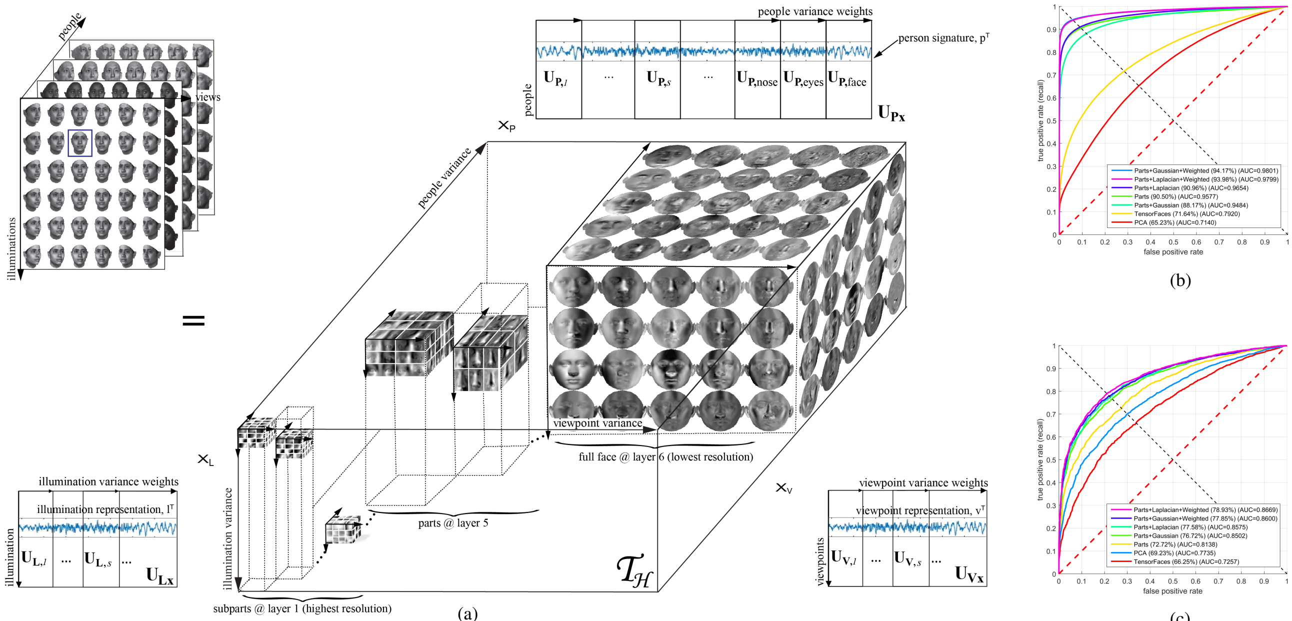

Figure 4: a) Compositional Hierarchical TensorFaces learns a hierarchy of features and represents each person as a part-based compositional representation. Figure depicts the training data factorization, $\mathcal{D} = \mathcal{T}_{\mathcal{H}} \times_L \mathbf{U}_L \times_V \mathbf{U}_V \times_P \mathbf{U}_P$, where an observation is represented as $\mathrm{d}(\mathrm{p}, \mathrm{v}, \mathrm{l}) = \mathcal{T}_{\mathcal{H}} \times_L \mathrm{l}^T \times_V \mathrm{v}^T \times_P \mathrm{p}^T$. b) ROC curves for the University of Freiburg 3D Morphable Faces dataset. c) ROC curves for the LFW dataset. The average accuracies are listed next to each method, along with the area under the curve (*AUC*). *Parts* refers to using Compositional Hierarchical TensorFaces models to separately analyze facial parts. *Gaussian*, *Laplacian* refers to using Compositional Hierarchical TensorFaces on a Gaussian/Laplacian data pyramid.

Test Dataset	**PCA**	**TensorFaces**	**Compositional Hierarchical TensorFaces**				
			Pixels	Gaussian Pyramid	Weighted Gaussian Pyramid	Laplacian Pyramid	Weighted Laplacian Pyramid
Freiburg	65.23%	71.64%	90.50%	88.17%	94.17%	90.96%	93.98%
LFW	69.23% ±1.51	66.25% ±1.60	72.72% ±2.14	76.72% ±1.65	77.85% ±1.83	77.58% ±1.45	78.93% ±1.77

Table 1: Empirical results reported for LFW : PCA, TensorFaces and Compositional Hierarchical TensorFaces. *Pixels* denotes indepent facial part analysis*Gaussian/Laplacian* use a multiresolution pyramid to analyze facial features at different scales. *Weighted* denotes a weighted nearest neighbhor.
Freiburg Experiment:
Train on Freiburg: 6 views (±60°,±30°,±5°); 6 illuminations (±60°,±30°,±5°), 45 people
Test on Freiburg: 9 views (±50°, ±40°, ±20°, ±10°, 0°), 9 illums (±50°, ±40°, ±20°, ±10°, 0°), 45 different people
LFW Experiment: Models were trained on approximately half of one percent ($0.5\% < 1\%$) **of the** 4.4**M images used to train DeepFace.**
Train on Freiburg:
15 views (±60°,±50°, ±40°,±30°, ±20°, ±10°,±5°, 0°), 15 illuminations (±60°,±50°, ±40°,±30°, ±20°, ±10°,±5°, 0°), 100 people
Test on LFW: We report the mean accuracy and standard deviation across standard literature partitions [30], following the *Unrestricted, labeled outside data* supervised protocol.

5 COMPOSITIONAL HIERARCHICAL TENSORFACES

Training Data: In our experiments, we employed gray-level facial training images rendered from 3D scans of 100 subjects. The scans were recorded using a CyberwareTM 3030PS laser scanner and are part of the 3D morphable faces database created at the University of Freiburg [6]. Each subject was combinatorially imaged in Maya from 15 different viewpoints ($\theta = -60°$ to $+60°$ in $10°$ steps on the horizontal plane, $\phi = 0°$) with 15 different illuminations ($\theta = -35°$ to $+35°$ in $5°$ increments on a plane inclined at $\phi = 45°$).

Data Preprocessing: Facial images were warped to an average face template by a piecewise affine transformation given a set of facial landmarks obtained by employing Dlib software [34, 36, 57]. Illumination was normalized with an adaptive contrast histogram equalization algorithm, but rather than performing contrast correction on the entire image, subtiles of the image were contrast normalized, and tiling artifacts were eliminated through interpolation. Histogram clipping was employed to avoid over-saturated regions. Each image, $\mathbf{d} \in \mathbb{R}^{I_0 \times 1}$, was convolved with a set of filters $\{\mathbf{H}_s \| s = 1 \ldots S\}$, and the filtered images, $\mathbf{d} \times_0 \mathbf{H}_s$, resulted either in a Gaussian or Laplacian image pyramid. Facial parts were segmented from the various layers.

Experiments: The composite person signature was computed for every test image by employing the multilinear projection algoritm[67, 73], and signatures were compared with weighted nearest neighbhor.

To validate the effectiveness of our system on real-world images, we report results on "LFW" dataset (*LFW*) [30]. This dataset contains 13,233 facial images of 5,749 people. The photos are unconstrained (*i.e.*, "in the wild"), and include variation due to pose, illumination, expression, and occlusion. The dataset consists of 10 train/test splits of the data. We report the mean accuracy and standard deviation across all splits in Table 1. Figure 4(b-c) depicts the experimental ROC curves. We follow the supervised *"Unrestricted, labeled outside data"* paradigm.

Results: While we cannot celebrate closing the gap on human performance, our results are promising. DeepFace, a CNN model, improved the prior art verification rates on LFW from 70% to 97.35%, by training on $4.4M$ images of 200×200 pixels from $4,030$ people, the same order of magnitude as the number of people in the LFW database. We trained on less than one percent (1%) of the 4.4M total images used to train DeepFace. Images were rendered from 3D scans of 100 subjects with an the intraocular distance of approximately 20 pixels and with a facial region captured by $10,414$ pixels (image size $\approx 100 \times 100$ pixels). We have currently achieved verification rates just shy of 80% on LFW. When data is limited, CNN models do not convergence or generalize.

Algorithm 2 Compositional Hierarchical Tensor Factorization.

Input: Data tensor, $\mathcal{D} \in \mathbb{R}^{I_0 \times I_1 \times \cdots \times I_C}$, filters $\mathbf{H}_s$, and desired dimensionality reduction $\tilde{J}_1, \ldots, \tilde{J}_C$.

1.*Initialization:*

1a. Decompose each data tensor segment, $\mathcal{D}_s = \mathcal{D} \times \mathbf{H}_s$, by employing the M-mode SVD.

$$\mathcal{D}_s = \mathcal{Z} \times_0 \mathbf{U}_{0,s} \cdots \times_c \mathbf{U}_{c,s} \cdots \times_C \mathbf{U}_{C,s}$$

1b. For $c = 0, 1, \ldots, C$, set $\mathbf{U}_{cx} = [\mathbf{U}_{c,1} \ldots \mathbf{U}_{c,s} \ldots \mathbf{U}_{c,S}]$, and truncate to $\tilde{J}_c$ columns by sorting all the eigenvalues from all data segments and deleting the columns corresponding to the lowest eigenvalues from various $\mathbf{U}_{c,s}$ and the rows from $\mathbf{Z}_{s[c]}$

2.*Optimization via alternating least squares:*

Iterate for $n := 1, \ldots, N$

For $c := 0, \ldots, C$,

2a. Compute mode matrix $\mathbf{U}_{cx}$ while holding the rest fixed.

$$\mathbf{U}_{cx} := \mathbf{D}_{[c]} \left(\mathbf{U}_{Cx} \odot \ldots \mathbf{U}_{(c+1)x} \odot \mathbf{U}_{(c-1)x} \odot \ldots \mathbf{U}_{0x} \right)^{\mathrm{T}+} \begin{bmatrix} \mathbf{Z}^{+}_{0[c]} & \mathbf{0} & & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} & & \vdots \\ \vdots & \mathbf{0} & \mathbf{Z}^{+}_{s[c]} & \mathbf{0} & \\ & & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & & \mathbf{0} & \mathbf{Z}^{+}_{S[c]} \end{bmatrix}$$

Set $\hat{\mathbf{U}}_{c,s}$ to the $\tilde{R}_{c,s}$ leading left-singular vectors of the SVD[a] of $\mathbf{U}_{c,s}$, a subset of the columns in $\mathbf{U}_{cx}$.Update $\mathbf{U}_{c,s}$ in $\mathbf{U}_{cx}$ with $\hat{\mathbf{U}}_{c,s}$

2b. Set the non-zero(nz) entries of $\mathrm{vec}(\mathcal{Z}_{\mathcal{H}})_{nz}$ based on:

$$\mathrm{vec}(\mathcal{Z}_{\mathcal{H}})_{nz} = (\mathbf{U}_{Cx} \otimes \cdots \otimes \mathbf{U}_{cx} \otimes \cdots \otimes \mathbf{U}_{0x})^{+}_{nz} \mathrm{vec}(\mathcal{D}).$$

until convergence. [b]

Output converged matrices $\mathbf{U}_{1x}, \ldots, \mathbf{U}_{Cx}$ and tensor $\mathcal{Z}_{\mathcal{H}}$.

[a] The complexity of computing the SVD of an $m \times n$ matrix $\mathbf{A}$ is $O(mn \min(m, n))$, which is costly when both m and n are large. However, we can efficiently compute the $\tilde{R}$ leading left-singular vectors of $\mathbf{A}$ by first computing the rank-$\tilde{R}$ modified Gram Schmidt (MGS) orthogonal decomposition $\mathbf{A} \approx \mathbf{QR}$, where $\mathbf{Q}$ is $m \times \tilde{R}$ and $\mathbf{R}$ is $\tilde{R} \times n$, and then computing the SVD of $\mathbf{R}$ and multiplying it as follows: $\mathbf{A} \approx \mathbf{Q}(\tilde{\mathbf{U}}\mathbf{S}\mathbf{V}^{\mathrm{T}}) = (\mathbf{Q}\tilde{\mathbf{U}})\mathbf{S}\mathbf{V}^{\mathrm{T}} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}}$.

[b] Note that N is a pre-specified maximum number of iterations. A possible convergence criterion is to compute at each iteration the approximation error $e_n := \|\mathcal{D} - \tilde{\mathcal{D}}\|^2$ and test if $e_{n-1} - e_n \leq \epsilon$ for sufficiently small tolerance ϵ.

6 CONCLUSION

In analogy to autoencoders which are inefficient neural network implementation of principal component analysis, a pattern analysis method based in linear algebra, CNNs are neural network implementations of tensor factorizations. This paper contributes to the tensor algebraic paradigm and models cause-and-effect as multilinear tensor interaction between intrinsic and extrinsic hierarchical causal factors of data formation. The data tensor is re-conceptualized into a hierarchical data tensor; a unified tensor model of wholes and parts is proposed; and a new compositional hierarchical tensor factorization is derived. Resulting causal factor representations are interpretable, hierarchical, and statistically invariant to all other causal factors. Our approach is demonstrated in the context of facial images by training on a very small set of synthetic images. While we have not closed the gap on human performance, we report encouraging face verification results on two test data sets–the Freiburg, and the Labeled Faces in the Wild datasets. CNN verification rates improved the 70% prior art to 97.35% when they employed 4.4M images from $4,030$ people, the same order of magnitude as the number of people in the LFW database. We have currently achieved verification rates just shy of 80% on LFW by employing synthetic images from 100 people for a total of less than one percent (1%) of the total images employed by DeepFace. By comparison, when data is limited, CNN models do not convergence, or generalize.

ACKNOWLEDGEMENT

Thank you to NYU Prof. Ernie Davis for feedback on this document.

REFERENCES

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] M. Bartlett, J. Movellan, and T. Sejnowski. Face recognition by independent component analysis. *IEEE Transactions on Neural Networks*, 13(6):1450–64, 2002.

[3] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. J. Goodfellow, A. Bergeron, N. Bouchard, and Y. Bengio. Theano: New features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS Workshop, 2012.

[4] P. N. Belhumeur, J. a. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(7):711–720, Jul 1997.

[5] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio. Theano: A CPU and GPU math expression compiler. In *Proc. Python for Scientific Computing Conf. (SciPy)*, Jun 2010. Oral Presentation.

[6] V. Blanz and T. A. Vetter. Morphable model for the synthesis of 3D faces. In *Proc. ACM SIGGRAPH 99 Conf.*, pages 187–194, 1999.

[7] R. Bro. Parafac: Tutorial and applications. In *Chemom. Intell. Lab Syst., Special Issue 2nd Internet Cont. in Chemometrics (INCINC'96)*, volume 38, pages 149–171, 1997.

[8] C. Cadieu and B. Olshausen. Learning transformational invariants from natural movies. In *Proc. 19th International Conf. on Neural Information Processing Systems*, NIPS'09, page 209216, 2009.

[9] J. D. Carroll and J. J. Chang. Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart-Young' decomposition. *Psychometrika*, 35:283–319, 1970.

[10] J. C. Chen, R. Ranjan, A. Kumar, C. H. Chen, V. M. Patel, and R. Chellappa. An end-to-end system for unconstrained face verification with deep convolutional

neural networks. In *IEEE International Conf. on Computer Vision Workshop (ICCVW)*, pages 360–368, Dec 2015.

[11] N. Cohen, O. Sharir, and A. Shashua. On the expressive power of deep learning: A tensor analysis. In *29th Annual Conf. on Learning Theory*, pages 698–728, 2016.

[12] N. Cohen and A. Shashua. Convolutional rectifier networks as generalized tensor decompositions. In *Proc. 33 rd International Conf. on Machine Learning*, 2016.

[13] R. Collobert, K. Kavukcuoglu, and C. Farabet. Torch7: A MATLAB-like environment for machine learning. In *BigLearn, NIPS Workshop*, number EPFL-CONF-192376, 2011.

[14] P. Common. Independent component analysis, a new concept? *Signal Processing*, 36:287–314, 1994.

[15] L. de Lathauwer. *Signal Processing Based on Multilinear Algebra*. PhD thesis, Katholieke Univ. Leuven, Belgium, 1997.

[16] L. de Lathauwer. Decompositions of a higher-order tensor in block termspart ii: Definitions and uniqueness. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1033–1066, 2008.

[17] L. de Lathauwer, B. de Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM Journal of Matrix Analysis and Applications*, 21(4):1253–78, 2000.

[18] L. de Lathauwer, B. de Moor, and J. Vandewalle. On the best rank-1 and rank-$(R_1, R_2, \ldots, R_n)$ approximation of higher-order tensors. *SIAM Journal of Matrix Analysis and Applications*, 21(4):1324–42, 2000.

[19] L. de Lathauwer and D. Nion. Decompositions of a higher-order tensor in block terms–part iii: Alternating least squares algorithms. *SIAM Journal on Matrix Analysis and Applications*, page 10671083, 2008.

[20] V. de Silva and L.-H. Lim. Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1084–1127, 2008.

[21] C. Figdor. Intrinsically/extrinsically. *Journal of Philosophy*, 105:691–718, 2008.

[22] H. Gao, H. K. Ekenel, M. Fischer, and R. Stiefelhagen. Multi-resolution local appearance-based face verification. In *Proc 20th International Conf. on Pattern Recognition (ICPR)*, pages 1501–04, Aug 2010.

[23] L. Grasedyck. Hierarchical singular value decomposition of tensors. *SIAM Journal on Matrix Analysis and Applications*, 31(4):2019–54, 2010.

[24] W. Hackbusch and S. Kühn. A new scheme for the tensor representation. *Journal of Fourier Analysis and Applications*, 15(5):706–722, 2009.

[25] R. Harshman. Foundations of the PARAFAC procedure: Model and conditions for an explanatory factor analysis. Technical Report Working Papers in Phonetics 16, University of California, Los Angeles, Los Angeles, CA, Dec 1970.

[26] B. Heisele, P. Ho, J. Wu, and T. A. Poggio. Face recognition: Component-based versus global approaches. *Computer Vision and Image Understanding*, 91:6–21, 2003.

[27] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural Comput.*, 18(7):1527–54, Jul 2006.

[28] E. Hsu, K. Pulli, and J. Popovic. Style translation for human motion. *ACM Transactions on Graphics*, 24(3):1082–89, 2005.

[29] G. B. Huang. Learning hierarchical representations for face verification with convolutional deep belief networks. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2518–25, Jun 2012.

[30] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, Oct 2007.

[31] I. Humberstone. Intrinsi/extrinsic. *Synthese*, 108:206–267, 1986.

[32] D. Hume. *An Enquiry Concerning Human Understanding, 2nd ed.* Andrew Millar, 1748.

[33] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition? In *Proc. International Conf. on Computer Vision (ICCV 2009)*, page 214653. IEEE, 2009.

[34] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, CVPR '14, pages 1867–74, Washington, DC, USA, 2014. IEEE Computer Society.

[35] Y. Kim, E. Park, S. Yoo, T. Choi, L. Yang, and D. Shin. Compression of deep convolutional neural networks for fast and low power mobile applications. *CoRR*, abs/1511.06530, 2015.

[36] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

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

[38] J. Kossaifi, A. Khanna, Z. C. Lipton, T. Furlanello, and A. Anandkumar. Tensor contraction layers for parsimonious deep nets. In *Computer Vision and Pattern Recognition (CVPR), Tensor Methods in Computer Vision Workshop*, pages 1940–46, Jul 2017.

[39] H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio. An empirical evaluation of deep architectures on problems with many factors of variation. In *Proc. 24th International Conf. on Machine Learning*, ICML '07, pages 473–480, New York, NY, USA, 2007. ACM.

[40] V. Lebedev, Y. Ganin, M. Rakhuba, I. V. Oseledets, and V. S. Lempitsky. Speeding-up convolutional neural networks using fine-tuned cp-decomposition. *CoRR*, abs/1412.6553, 2014.

[41] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998.

[42] D. Lewis. Extrinsic properties. *Philosophical Studies*, 44:197–200, 1983.

[43] D. K. Lewis. Causation. *Journal of Philosophy*, pages 556–567, 1973.

[44] H. Li and G. Hua. Hierarchical-pep model for real-world face recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 4055–64, Jun 2015.

[45] A. M. Martinez and A. C. Kak. PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):228–233, Feb 2001.

[46] P. Menzies. Intrinsic vs. extrinsic conceptions of causation. In H. Sankey, editor, *Causation and the Laws of Nature*, pages 313–29. 1999.

[47] K. Murphy, A. Torralba, D. Eaton, and W. Freeman. *Object Detection and Localization Using Local and Global Features*, pages 382–400. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[48] V. Nair and G. E. Hinton. 3d object recognition with deep belief nets. In *Proc. 22Nd International Conf. on Neural Information Processing Systems*, NIPS'09, pages 1339–47, USA, 2009. Curran Associates Inc.

[49] A. Novikov, D. Podoprikhin, A. Osokin, and D. P. Vetrov. Tensorizing neural networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 442–450. Curran Associates, Inc., 2015.

[50] I. V. Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317, 2011.

[51] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In M. W. J. Xianghua Xie and G. K. L. Tam, editors, *Proc. British Machine Vision Conference (BMVC)*, pages 41.1–41.12. BMVA Press, September 2015.

[52] J. Pearl and C. U. Press. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[53] I. Perros, R. Chen, R. Vuduc, and J. Sun. Sparse hierarchical tucker factorization and its application to healthcare. In *Proc. IEEE International Conf. on Data Mining*, pages 943–948, Nov 2015.

[54] M. Ranzato, C. Poultney, S. Chopra, and Y. LeCun. Efficient learning of sparse representations with an energy-based model. In *Proc. 19th International Conf. on Neural Information Processing Systems*, NIPS'06, pages 1137–44, Cambridge, MA, USA, 2006. MIT Press.

[55] C. R. Rao and S. K. Mitra. *Generalized Inverse of Matrices and its Applications*. Wiley, New York, 1971.

[56] G. Shakhnarovich and B. Moghaddam. Face recognition in subspaces. In S. Z. Li and A. K. Jain, editors, *Handbook of Face Recognition*, pages 141–168. Springer, New York, 2004.

[57] W. Si, K. Yamaguchi, and M. A. O. Vasilescu. *Face Tracking with Multilinear (Tensor) Active Appearance Models*. Jun 2013. http://pdfs.semanticscholar.org/6c64/59d7cadaa210e3310f3167dc181824fb1bff.pdf.

[58] N. D. Sidiropoulos, L. D. Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos. Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing*, 65:3551–3582, 2017.

[59] Y. Sun, X. Wang, and X. Tang. Hybrid deep learning for face verification. In *Proc. IEEE International Conf. on Computer Vision (ICCV)*, pages 1489–96, Dec 2013.

[60] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proc.IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1701–08, Jun 2014.

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

[62] M. A. Turk and A. P. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[63] M. A. O. Vasilescu. An algorithm for extracting human motion signatures. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Hawai, 2001.

[64] M. A. O. Vasilescu. Human motion signatures for character animation. In *ACM SIGGRAPH 2001 Conf. Abstracts and Applications*, Los Angeles, Aug 2001.

[65] M. A. O. Vasilescu. Human motion signatures: Analysis, synthesis, recognition. In *Proc. Int. Conf. on Pattern Recognition*, volume 3, pages 456–460, Quebec City, Aug 2002.

[66] M. A. O. Vasilescu. *A Multilinear (Tensor) Algebraic Framework for Computer Graphics, Computer Vision, and Machine Learning*. PhD thesis, University of Toronto, Department of Computer Science, 2009.

[67] M. A. O. Vasilescu. Multilinear projection for face recognition via canonical decomposition. In *Proc. IEEE International Conf. on Automatic Face Gesture Recognition and Workshops (FG 2011)*, pages 476–483, Mar 2011.

[68] M. A. O. Vasilescu and D. Terzopoulos. Multilinear analysis for facial image recognition. In *Proc. Int. Conf. on Pattern Recognition*, volume 2, pages 511–514, Quebec City, Aug 2002.

[69] M. A. O. Vasilescu and D. Terzopoulos. Multilinear analysis of image ensembles: TensorFaces. In *Proc. European Conf. on Computer Vision (ECCV 2002)*, pages 447–460, Copenhagen, Denmark, May 2002.

[70] M. A. O. Vasilescu and D. Terzopoulos. Multilinear subspace analysis of image ensembles. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume II, pages 93–99, Madison, WI, 2003.

[71] M. A. O. Vasilescu and D. Terzopoulos. TensorTextures: Multilinear image-based rendering. *ACM Transactions on Graphics*, 23(3):336–342, Aug 2004. Proc. ACM SIGGRAPH 2004 Conf., Los Angeles, CA.

[72] M. A. O. Vasilescu and D. Terzopoulos. Multilinear independent components analysis. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume I, pages 547–553, San Diego, CA, 2005.

[73] M. A. O. Vasilescu and D. Terzopoulos. Multilinear projection for appearance-based recognition in the tensor framework. In *Proc. 11th IEEE International Conf. on Computer Vision (ICCV'07)*, pages 1–8, Rio de Janeiro, Brazil, 2007.

[74] D. Vlasic, M. Brand, H. Pfister, and J. Popovic. Face transfer with multilinear models. *ACM Transactions on Graphics (TOG)*, 24(3):426–433, Jul 2005.

[75] Y. Wong, M. T. Harandi, C. Sanderson, and B. C. Lovell. On robust biometric identity verification via sparse encoding of faces: Holistic vs local approaches. In *Proc. International Joint Conf. on Neural Networks (IJCNN)*, pages 1–8, June 2012.

[76] C. Xiong, L. Liu, X. Zhao, S. Yan, and T. K. Kim. Convolutional fusion network for face verification in the wild. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(3):517–528, Mar 2016.

[77] Z. Xu, H. Chen, S. Zhu, and J. Luo. A composite template for human face modeling and sketch. *IEEE Transations on Pattern Analysis and Machine Intelligence*, 30:955–969, 2008.

[78] M.-H. Yang. Face recognition using kernel methods. In *Proc. 14th International Conf. on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, pages 1457–64, Cambridge, MA, USA, 2001. MIT Press.