Analysis of Temporal Tensor Datasets on Product Grassmann Manifold

Bojan Batalo
University of Tsukuba
Tsukuba, Japan
bojan_batalo@cvlab.cs.tsukuba.ac.jp

Lincon S. Souza
AIST
Tokyo, Japan
lincon.souza@aist.go.jp

Bernardo B. Gatto
AIST
Tsukuba, Japan
bernardo.gatto@aist.go.jp

Naoya Sogi
University of Tsukuba
Tsukuba, Japan
sogi@cvlab.cs.tsukuba.ac.jp

Kazuhiro Fukui
University of Tsukuba
Tsukuba, Japan
kfukui@cs.tsukuba.ac.jp

Abstract

Growing abundance of multi-dimensional data creates a need for efficient data exploration and analysis. In this paper, we address this need by tackling the task of tensor dataset visualization and clustering, as tensors are a natural form of multi-dimensional data. Previous work has shown that representing individual tensor modes via respective linear subspaces and unifying them on the product Grassmann manifold (PGM) is an effective and memory-efficient way of representation. However, such representation may lead to loss of valuable temporal information. To address this issue, we model temporal tensor modes with a Hankel-like matrix, preserving sequence information and encoding it with a linear subspace, fully compatible with PGM. Unifying regular tensor modes and Hankel-like representation of regular tensor modes then enriches representation on the PGM, with minimal increase in computational complexity. By relying on geodesic distance on the manifold, we facilitate analysis of multi-dimensional datasets in two ways: 1) by enabling straightforward visualizations using algorithms such as t-SNE; and 2) by fostering clustering of data using distance- or similarity-based methods such as spectral clustering. We evaluate our approach on hand gesture and action recognition datasets as exemplars of temporal tensor datasets.

1. Introduction

Data exploration and representation has been at the forefront of problems tackled by the pattern recognition community for a long time. Increase of multi-dimensional data sources, such as various sensors and cameras, has been followed by a steady rise of representation techniques. Their common aim is to naturally mold and express rich information contained in the data in a unified way. Such representation then enables analysis via visualisations, clustering and classification, allowing for a deeper insight into the structure of a given dataset. In this paper we consider representation of multi-dimensional data with temporal information, and their visualization and clustering as first steps of data exploration.

A representation method of particular interest is the *tensorial* form, where the nature of multi-dimensional data is preserved without changing its inherent structure. Concretely, a single data point is represented as a single tensor. In addition to the simplicity of singular representation, this approach brings the benefits of established multilinear algebra, allowing for efficient computations on tensor data. Applications such as hand gesture and action recognition [5, 15, 22], medical data analysis and imaging [16, 27, 28, 31, 42, 46], multi-spectral imaging [3, 17–19, 23] and others [4, 32] may benefit from tensor representation.

One way to efficiently represent a tensor as a single data point is representation on the product Grassmann manifold (PGM) [26], rooted in tensor unfolding [21] and subspace representation [6, 30, 43]. An n-dimensional tensor can be unfolded along each of its dimensions, generating n modes [21]. With each mode essentially presenting a unique look into the data contained within the tensor, by analyzing separate modes it is possible to extract information inaccessible by other means. For example, a video can be viewed as a 3-dimensional tensor containing two spatial and one temporal mode, and each mode can be compactly represented by a respective linear subspace it spans.

A Grassmann manifold (GM) is defined as a set of linear subspaces of same dimension, and can be geometrically interpreted as a surface where these subspaces are points on the manifold. Therefore, there is a GM corresponding to each of the tensor modes. A single manifold expresses geo-

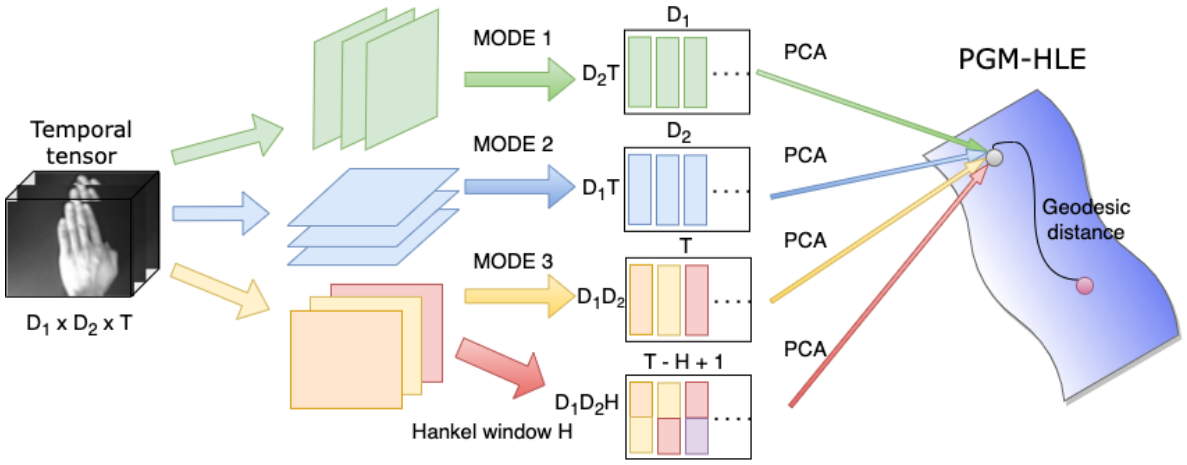

Figure 1. A temporal tensor can be unfolded along dimensions D_1, D_2 and T to generate mode features. Hankel-like matrix is created from temporal mode by applying a sliding window of size H to preserve sequential information. Unified tensor representation is created on product Grassmann manifold with Hankel-like embedding (PGM-HLE).

metrical relations between subspaces of the same mode via geodesic distance, enabling discriminative analysis within the manifold [12, 14, 34, 36, 37, 41].

However, utilizing information from each mode in a unified manner requires the construction of a PGM from distinct factor (mode) manifolds [7, 25]. Analysis can then be performed on the PGM in a similar vein, taking into account that chordal distance between points on PGM is equivalent to the Cartesian product of geodesic distances on respective mode Grassmannians [13, 25]. The downside is that representing all tensor modes in the same way can lead to loss of mode-specific information. For videos, this means potential loss of discriminative temporal features, as linear subspaces do not fully preserve sequence information [9].

In this paper, we address the lack of explicit handling of temporal information, by exploiting the fact that subspace representations of data on any number of factor manifolds can be used on the PGM. To this end, we model the temporal mode with a Hankel-like matrix, which can then be encoded with a sequence-preserving linear subspace and incorporated with regular tensor mode representations via PGM. We refer to this representation as *product Grassmann Manifold with Hankel-like embedding* (PGM-HLE), shown on Fig. 1. The idea of Hankel-like representations is strongly motivated by its recent applications for sequential data, as investigated in the literature [8, 24, 33, 35, 39].

Further analysis of tensors on PGM-HLE is done in the context of chordal distance between two points on the manifold [11, 38], a metric that unifies distances between subspaces within a single factor Grassmannian [2]. In this way, we provide valuable temporal context at a minimal increase in computational complexity. Besides, our approach demonstrates usability of specialized representations via PGM.

Our main contributions are summarized as follows:

1. We introduce product Grassmann Manifold with Hankel-like embedding (PGM-HLE), a temporal tensor representation method based on tensor decomposition and Hankel-like matrix.

2. Next, we show the benefit of using PGM-HLE through visualizations. In addition, we demonstrate a simple use-case of tensor dataset visualization with t-SNE and PGM-HLE.

3. Finally, we evaluate spectral clustering on PGM-HLE with hand gesture and action recognition datasets.

The rest of the paper is organized as follows. In Sec. 2 we describe PGM-HLE in detail, and how to use it for t-SNE visualization and clustering. Sec. 3 describes the datasets and experimental results. Finally, with Sec. 4 we conclude the paper and suggest future work.

2. PROPOSED METHOD

In this section, we describe the proposed tensor representation on PGM-HLE, shown on Fig. 2. First, we formulate the problem of tensor representation. Next, explain the n-mode tensor representation of spatio-temporal features via linear subspaces and then introduce Hankel-like embedding of temporal modes. We describe our full tensor representation on the PGM with Hankel-like embedding. Finally, we show how PGM-HLE enables distance and similarity-based visualization and clustering algorithms such as t-SNE and spectral clustering (SC).

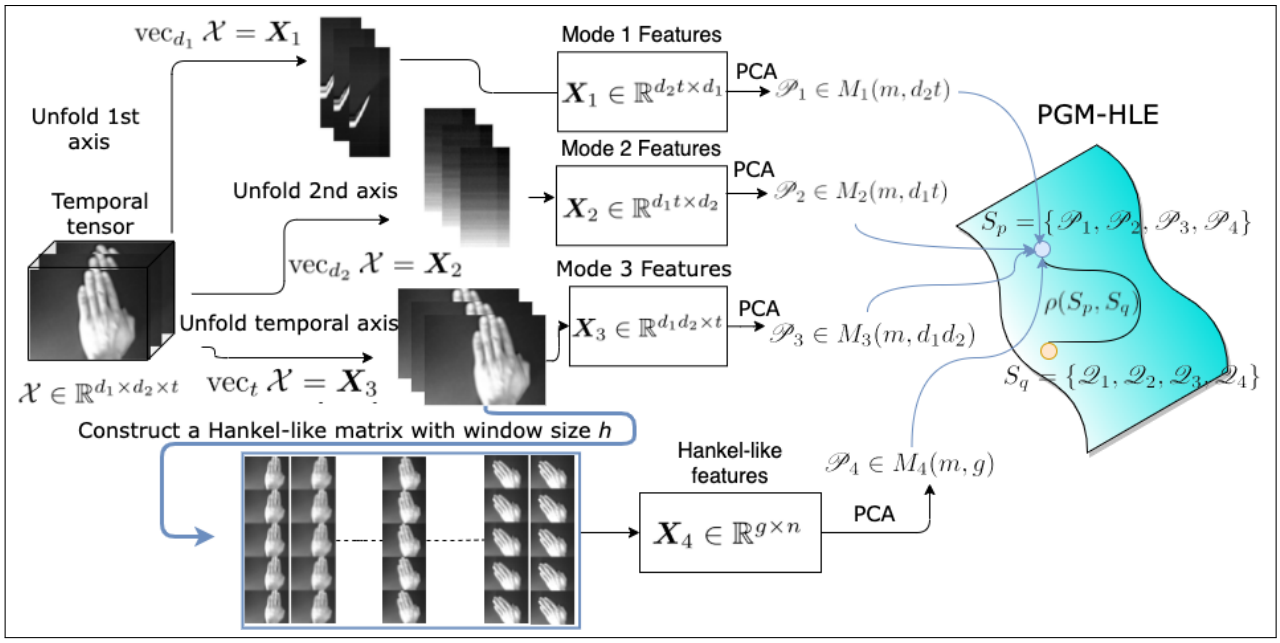

Figure 2. A greyscale tensor $\mathcal{X}$ is unfolded to generate mode features $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3$, and Hankel-like matrix $\boldsymbol{X}_4$. non-centered PCA is performed to generate a set of subspace basis vectors. Tensors $\mathcal{X}$ and $\mathcal{Y}$, represented by sets of subspaces S_p and S_q are compared using geodesic distance $\rho(S_p, S_q)$.

Notation is as follows. Scalars are denoted by lowercase letters and sets are denoted by uppercase letters. Vectors and matrices are denoted by boldface lowercase and uppercase letters respectively. Calligraphic letters denote tensors and script letters denote subspaces. Given a matrix $\boldsymbol{A} \in \mathbb{R}^{w \times h}$, $\boldsymbol{A}^\top \in \mathbb{R}^{h \times w}$ denotes its transpose.

2.1. Basic Idea

For simplicity, in this work, multi-dimensional data points are regarded as 3-mode tensors $\mathcal{X}$ of size $d_1 \times d_2 \times t$, where the mode of size t is a temporal mode. However, the proposed approach can be generalized to temporal tensors with n modes. In its raw form, the data can be rather unwieldy and uninformative. Therefore we compactly represent a tensor $\mathcal{X}$ with a set of *mode subspaces*. This representation has multiple advantages: it allows parallel processing, and helps finding correlations among various factors inherent in each mode.

We formulate the tensor representation problem as follows: Let $X = \{\mathcal{X}_i\}_{i=1}^n$ be a set of n tensors. In addition, let $T(\mathcal{X})$ be a transformation of a tensor in its raw form to its representation on PGM-HLE. Finally let $\rho(\mathcal{X}, \mathcal{Y})$ be the similarity function between two tensors $\mathcal{X}$ and $\mathcal{Y}$ defined by geometric properties of PGM-HLE.

We consider optimization problems where we minimize a function F dependent on some sort of distance $d(\mathcal{X}, \mathcal{Y})$ between tensors $\mathcal{X}$ and $\mathcal{Y}$. This can be written as:

$$\min F(d(\mathcal{X}, \mathcal{Y})), \tag{1}$$

where $\mathcal{X}, \mathcal{Y} \in X = \{\mathcal{X}_i\}_{i=1}^n$. We aim to create a transformation $T(\mathcal{X})$ which provides similarity $\rho(\mathcal{X}, \mathcal{Y})$ as an interface for solving Eq. (1) with $d(\mathcal{X}, \mathcal{Y}) = 1 - \rho(\mathcal{X}, \mathcal{Y})$.

2.2. n-mode Tensor Representation with Linear Subspaces

Let $\mathcal{X} \in \mathbb{R}^{d_1 \times d_2 \times t}$ be a 3-dimensional tensor, where t represents the temporal dimension. Tensor $\mathcal{X}$ is unfolded into a set of matrices $X = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3\}$, a process known as *matricization*. Consider a video as a tensor of size $t \times h \times w$, where w, h and t are width, height and number of frames, respectively. This tensor is unfolded into $X = \{\boldsymbol{X}_1 \in \mathbb{R}^{(wt) \times h}, \boldsymbol{X}_2 \in \mathbb{R}^{(ht) \times w}, \boldsymbol{X}_3 \in \mathbb{R}^{(wh) \times t}\}$, with each mode representing concatenated slices along a specific tensor dimension. Therefore, $\mathcal{X}$ can be decomposed to achieve a compact subspace representation using *n-mode SVD* [21], defined as:

$$\mathcal{X} = \mathcal{C} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 \times_3 \boldsymbol{U}_3. \tag{2}$$

Core tensor $\boldsymbol{C}$ contains values analogous to eigenvalues of *SVD*, while matrices $\{\boldsymbol{U}_j\}_{j=1}^{n=3}$ contain the singular vectors for each unfolded matrix $\boldsymbol{X}_j$, and expression $\boldsymbol{U}_j \boldsymbol{\Lambda}_j \boldsymbol{U}_j^\top = \boldsymbol{X}_j \boldsymbol{X}_j^T$ holds.

Subsequently, we select a subspace spanned by eigenvectors $\boldsymbol{U}_j$, resulting in a set of subspaces $S_p = \{\mathscr{P}_j\}_j^{n=3}$, spanned by basis vectors $\{\boldsymbol{P}_j\}_j^{n=3}$, where $\boldsymbol{P}_j \in \mathbb{R}^{f_j \times m_j}$, containing m_j eigenvectors corresponding to the highest m_j eigenvalues. Different m_j can be selected, as each mode exhibits different properties and levels of information density. In summary, tensor $\mathcal{X}$ is mapped to a set of mode subspaces S_p. This offers a compact representation and an opportunity to analyze each tensor mode independently. However, as *SVD* does not preserve sequence information, temporal features may be lost.

2.3. Hankel-like Embedding of Temporal Modes

It is possible to preserve sequential features from temporal tensor modes within a linear subspace. Consider the temporal mode of tensor $\mathcal{X} \in \mathbb{R}^{d_1 \times d_2 \times t}$:

$$\text{vec}_t\, \mathcal{X} = \boldsymbol{X}_t = \boldsymbol{X}_3 = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t]. \tag{3}$$

A *Hankel-like* matrix $\boldsymbol{H}$ can be created by applying a sliding window of size h over the columns of matrix $\boldsymbol{X}_3$. This creates a set of n lagged frame sequences comprised of h frames, arranged into matrix $\boldsymbol{H} \in \mathbb{R}^{g \times n}$ as follows:

$$\boldsymbol{H} = \begin{bmatrix} x_1 & x_2 & \ldots & x_n \\ x_2 & x_3 & \ldots & x_{n+1} \\ \vdots & & \ddots & \vdots \\ x_h & x_{h+1} & \ldots & x_{h+n-1} \end{bmatrix}. \tag{4}$$

The number of columns of matrix $\boldsymbol{H}$ is determined by the total length of sequence $[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t]$, given by relationship $n = t - h + 1$, and the number of rows is $g = h \times f_t$. In our method, h is considered a hyperparameter.

In temporal mode $\boldsymbol{X}_3$, the ordering of columns carries sequence information, which is lost when applying *SVD* to create a subspace representation. However, H preserves temporal information by embedding it in its columns. We then construct a compact subspace representation in the following manner:

$$\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^\top = \boldsymbol{H}\boldsymbol{H}^\top. \tag{5}$$

$\boldsymbol{V} \in \mathbb{R}^{g \times n}$ contains eigenvectors as columns and $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ is a diagonal matrix with n eigenvalues. Basis vectors of a subspace are obtained by selecting m eigenvectors $\boldsymbol{P}_4 = [\boldsymbol{v}_1, ..., \boldsymbol{v}_m]$ corresponding to the m highest eigenvalues. Resulting subspace $\mathscr{Q}_4$ spanned by $\boldsymbol{P}_4$ exhibits sequence-preserving qualities in each of the eigenvectors, which are generalized to the whole subspace. We call this representation Hankel-like embedding (HLE). Additionally, HLE fully is compatible with subspaces modeling spatio-temporal data from Sec. 2.2.

2.4. Product Grassmann Manifold with Hankel-like Embedding

By using the two approaches described in subsections 2.2 and 2.3, a temporal tensor $\mathcal{X} \in \mathbb{R}^{d_1 \times d_2 \times t}$ is represented by a set of subspaces $S_p = \{\mathscr{P}_j\}_{j=1}^4$, containing mode subspaces $\mathscr{P}_1$, $\mathscr{P}_2$ and $\mathscr{P}_3$ and the HLE subspace $\mathscr{P}_4$. Every $\mathscr{P}_j$ is a point on a Grassmann manifold $M_j(m_j, d_j)$, where m_j and d_j are dimensions of subspace $\mathscr{P}_j$ and feature space respectively. A unified representation is constructed on product Grassmann manifold from a set of factor manifolds $M = \{M_j\}_{j=1}^4$ as follows:

$$M = M_1 \times M_2 \times M_3 \times M_t = (\mathscr{P}_1, \mathscr{P}_2, \mathscr{P}_3, \mathscr{P}_4), \tag{6}$$

where $\times$ denotes Cartesian product. Therefore, each temporal tensor is represented as a single point on M. Further data analysis is possible using a metric defined on PGM, namely the geodesic distance between two points on the PGM. This is a natural choice of dissimilarity due to its utilization of the manifold surface [1].

To define geodesic distance on M, we first define geodesic distances between points (subspaces) on factor manifolds. These distances are parametrized in terms of canonical angles between subspaces, defined as minimal angles between two subspaces [6]. Given subspaces $\mathscr{P}$ and $\mathscr{Q}$ and their basis vectors $\boldsymbol{P}$ and $\boldsymbol{Q}$, canonical angles $\{0 \leq \theta_1, ..., \theta_m \leq \frac{\pi}{2}\}$ can be computed by SVD as:

$$\boldsymbol{P}^\top \boldsymbol{Q} = \boldsymbol{U}_p \boldsymbol{\Sigma} \boldsymbol{U}_q. \tag{7}$$

$\boldsymbol{U}_p$ and $\boldsymbol{U}_q$ contain canonical vectors, and $\boldsymbol{\Sigma} = \text{diag}(\kappa_1, \ldots, \kappa_r)$ is a diagonal matrix with m singular values $\{\kappa_l\}_{l=1}^m$. Canonical angles $\{\theta_l\}_{l=1}^m$ can be obtained as $\{\cos^- 1(\kappa_l)\}_{l=1}^m$. Similarity between subspaces $\mathscr{P}$ and $\mathscr{Q}$ is then defined as:

$$s(\mathscr{P}, \mathscr{Q}) = \frac{1}{m} \sum_{l=1}^{m} \cos^2 \theta_l. \tag{8}$$

Using this similarity in each factor manifold, we define the similarity on PGM. Tensors $\mathcal{X}$ and $\mathcal{Y}$ represented by sets of subspaces $\{\mathscr{P}_j\}_{j=1}^4$ and $\{\mathscr{Q}_j\}_{j=1}^4$ is defined as:

$$\rho(\mathcal{X}, \mathcal{Y}) = \frac{1}{n} \sqrt{\sum_{j=1}^{n=4} s(\mathscr{P}_j, \mathscr{Q}_j)^2}. \tag{9}$$

As individual similarities $s(\mathscr{P}_j, \mathscr{Q}_j)$ are bounded between 0 and 1, division by number of factor manifolds n is introduced to maintain same bounds for final geodesic distance. This enables the conversion of similarity to distance as $d(\mathcal{X}, \mathcal{Y}) = 1 - \rho(\mathcal{X}, \mathcal{Y})$. Having defined the similarity between points on PGM, it is possible to conduct further analysis of tensor datasets by considering their layout in the

manifold space. We explore an example of dataset visualization and clustering in Sec. 2.5 and Sec. 2.6 respectively.

2.5. Multilinear t-SNE

t-SNE [40] is a well known data visualization algorithm able effectively map multi-dimensional data to two or three dimensions, while preserving both global and local structure. This is achieved by representing distances between data points as probabilities under Gaussian distribution, and minimizing Kullback-Leibler divergence between joint probability distributions A and B in high-dimensional and low-dimensional spaces respectively. The following Eq. (10) represents this cost function and is optimized using gradient descent:

$$C = KL(A||B) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}. \tag{10}$$

Consider two data points x_i and x_j in high-dimensional space, and their low-dimensional mappings y_i and y_j. Here, p_{ij} is the probability of choosing x_j as a closely-related neighbour of x_i under a Gaussian distribution centered on x_i. Analogously, q_{ij} is the same probability with respect to y_i and y_j. Probability p_{ij}, given by the following equation:

$$p_{ij} = \frac{\exp(-||x_i - x_j||^2 / 2\sigma^2)}{\sum_{k \neq l} \exp(-||x_k - x_l||^2 / 2\sigma^2)}, \tag{11}$$

is calculated explicitly as part of t-SNE algorithm. It assumes Euclidean distance between points x_i and x_j, which is defined for multi-dimensional vectors. However, if data points x_i and x_j are instead multi-linear tensors $\mathcal{X}_i$ and $\mathcal{X}_j$, the Euclidean distance is not defined, and Eq. (11) cannot be solved. Therefore, t-SNE cannot be utilized to visualise temporal tensor datasets.

We have provided an interface to solve this problem in Sec. 2.4, by introducing a similarity function between two tensors based on the PGM with Eq. (9). Thus, we modify Eq. (11) in the following manner:

$$p_{ij} = \frac{\exp(-d(\mathcal{X}_i, \mathcal{X}_j) / 2\sigma^2)}{\sum_{k \neq l} \exp(-d(\mathcal{X}_k, \mathcal{X}_l) / 2\sigma^2)}, \tag{12}$$

where $d(\mathcal{X}, \mathcal{Y}) = 1 - \rho(\mathcal{X}, \mathcal{Y})$. By using Eq. (12) it is possible to optimize the Kullback-Leibler divergence as defined in Eq. (10) for datasets with temporal tensor datasets and make appropriate visualizations with t-SNE.

2.6. Clustering based on distance and affinity matrices

A variety of clustering algorithms relies on constructing symmetric matrices based on pairwise distances or similarities between all data points in the dataset. One notable example is spectral clustering [29]. Given a set of points

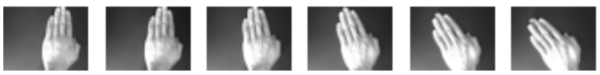

Figure 3. Uniformly sampled frames of temporal mode from CMB dataset.

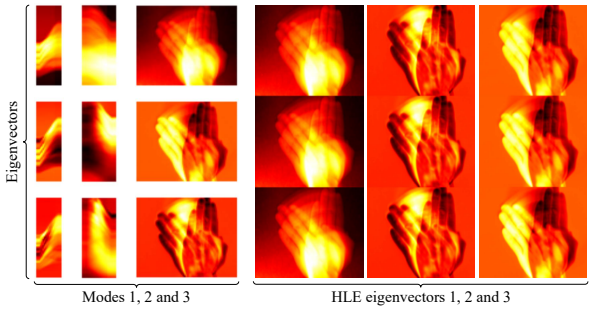

Figure 4. Subspace basis vectors of tensor modes 1, 2, 3 and Hankel-like embedding, exhibiting different form and information.

$S = \{s_1, s_2, \ldots, s_n\} \in \mathbb{R}^l$, the first step in spectral clustering is to construct an affinity matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, where $\boldsymbol{A}_{ij} = \exp(-||s_i - s_j||^2 / 2\sigma^2)$, $i \neq j$ and $A_{ii} = 0$. However, if we consider a set of multilinear tensors $X = \{\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n\}$ as a set of data points, the similarity defined in Eq. (9) can be used to rewrite the formation of affinity matrix to:

$$\boldsymbol{A}_{ij} = \rho(\mathcal{X}_i, \mathcal{X}_j). \tag{13}$$

Having constructed the affinity matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ using Eq. (13), spectral clustering algorithm proceeds with standard steps, without modifications, as defined in [29]. A major advantage of PGM-HLE representation is that via Eq. (9) it provides an interface for using various algorithms on tensor datasets, otherwise defined only for vector-valued data.

3. EXPERIMENTAL RESULTS

3.1. Datasets

We use **Cambridge Hand Gesture** [20] (CMB) and **UT-Kinect** [45] (UT) datasets in visualization and clustering experiments. **CMB** is a benchmark dataset for hand gesture recognition, and comprises 9 classes with 100 videos per class. It is divided equally into five sets based on illumination settings. We perform clustering experiments on each set, and on the entire dataset to evaluate robustness against illumination change. **UT** contains skeleton data of 10 action classes, performed by 10 subjects in front of a Kinect device. There are 200 sequences in total. The data contains x, y and z coordinates of 20 skeleton joints.

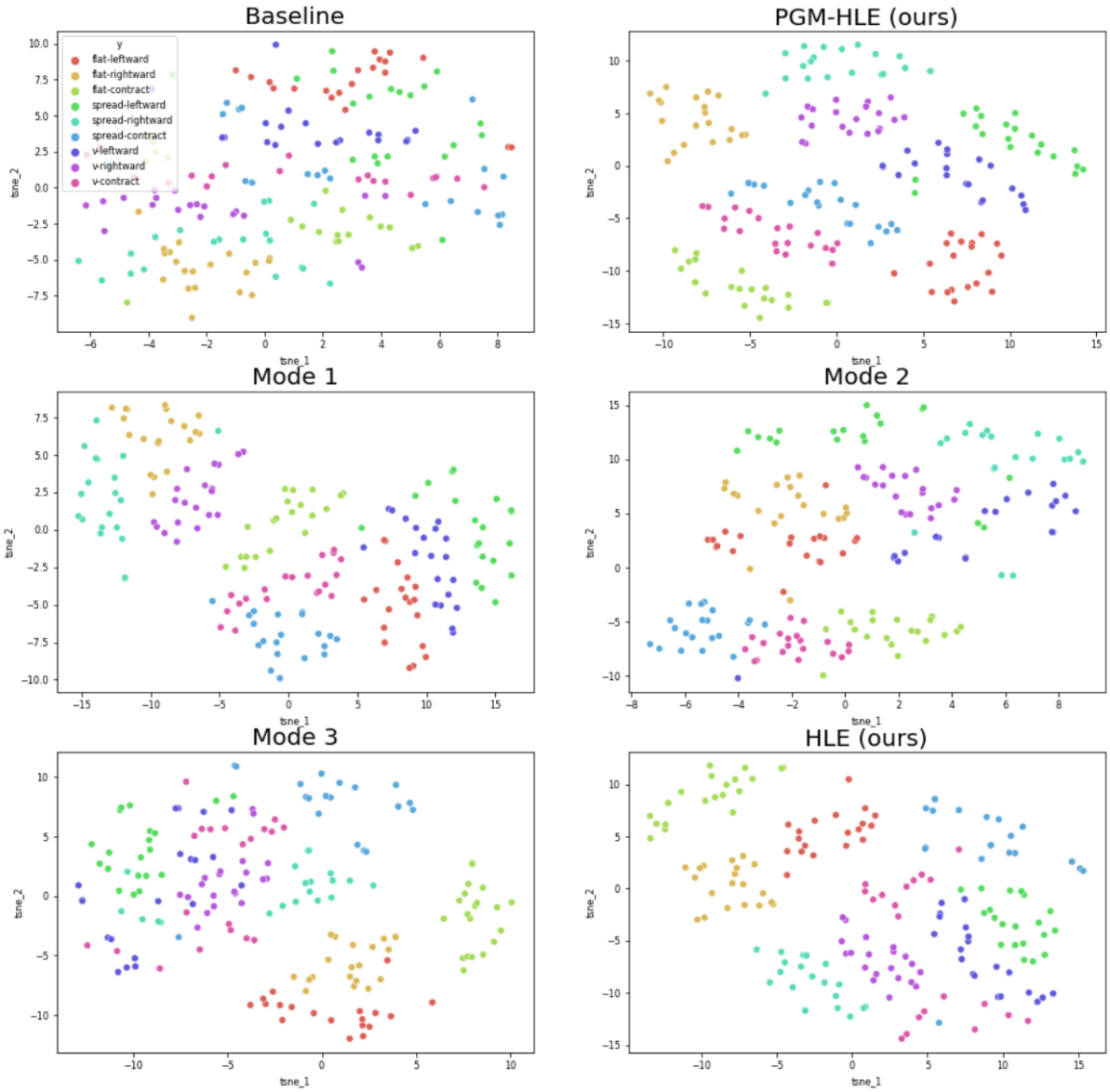

Figure 5. t-SNE visualizations of baseline vectorized representation, regular tensor modes, HLE and PGM-HLE for Cambridge Hand Gesture dataset. This dataset combines three hand shapes - 'flat', 'spread' and 'v' with three movements - 'leftward', 'rightward' and 'contract'. Baseline is very cluttered. Modes and HLE offer different perspectives and provide high discrimination among the clusters. PGM-HLE has the best defined visualization obtained by unifying those perspectives.

Tensors in **CMB** dataset are of shape $h \times w \times t$, where h, w and t stand for height, width and sequence length respectively, while in **UT** they are of shape $c \times j \times t$, where c, j and t stand for coordinate, joint and sequence length respectively. As all tensors have variable temporal dimension t, we resize them to $12 \times 16 \times 30$ (CMB) and $3 \times 20 \times 20$ (UT). We use video data as it is easy to visualize and interpret subspace basis vectors, as well as evaluate formed clusters.

3.2. Visualizations of Hankel-like subspaces

As videos are simple to interpret, we can visualize eigenvectors spanning subspaces of tensor modes, including the Hankel-like embedding (HLE) of temporal mode. This provides insight into information contained within these representations, as well as grounds for further interpretability. An example tensor $\mathcal{X}$ from **CMB** dataset, a closed hand moving to the left, is shown on Fig. 3.

We then decompose tensor $\mathcal{X}$ into three modes, obtain

Dataset	Baseline	M1	M2	M3	HLE	PGM	PGM-HLE
CMB S1	32.94%	88.88%	87.77%	84.44%	*95.00%*	97.22%	**98.33%**
CMB S2	16.88%	74.44%	80.00%	78.88%	*83.88%*	86.11%	**88.33%**
CMB S3	21.44%	74.44%	70.55%	72.77%	*78.88%*	81.11%	**82.22%**
CMB S4	27.55%	74.44%	71.11%	63.33%	*76.66%*	80.00%	**83.88%**
CMB S5	28.61%	78.88%	81.66%	80.55%	*84.44%*	86.66%	**88.33%**
Cambridge	18.80%	75.55%	69.77%	75.44%	*83.11%*	84.88%	**86.66%**
UT-Kinect	61.30%	*80.90%*	62.31%	65.32%	77.38%	92.46%	**93.96%**

Table 1. Spectral clustering results on CMB and UT. All tensor modes outperform the baseline. HLE works better than regular modes 1, 2 and 3, except on UT. PGM-HLE outperforms all baselines with 86.66% (CMB) and 93.96% (UT).

a Hankel-like matrix, perform non-centered PCA and construct respective subspaces. On Fig. 4 we depict first three eigenvectors of each mode. It can be clearly seen that each mode carries different information, with modes 1 and 2 being difficult to interpret. Eigenvectors of mode 3 are very similar to eigenvectors of Hankel-like subspace, with the latter being almost a concatenation of the former. However, in Sec. 3.4 we show the effect of this information on clustering accuracy.

3.3. Tensor visualization on TS-PGM

To investigate the effectiveness of PGM-HLE on **CMB** dataset, we use t-SNE on 1) baseline vectorized representation of tensors, 2) subspaces of individual modes, including the HLE and 3) on the PGM-HLE, and compare these visualizations. As t-SNE is not a deterministic algorithm, we run it 10 times and pick the one with lowest KL value [44]. Results are depicted on Fig. 5.

Baseline setting 1) results in the worst visualization, with a high KL divergence score of 0.371, while setting 3) achieves the best, with the lowest KL score of 0.208. Modes 1-4 in setting 2) showcase that each mode carries information of different characteristics and quality with respect to separability, with respective KL scores of 0.213, 0.271, 0.282 and 0.242. For example, mode 2 seems capable of separating all classes of 'contract' shape, mode 1 successfully extracts rightward movements and mode 3 'flat' hand shapes.

HLE appears very similar to temporal mode 3, which is expected due to underlying temporal information of both representations. However, separability between clusters in HLE is much higher and easier to notice. For example, HLE is able to group samples of classes 'v-rightward', 'v-leftward' and 'spread-rightward', while improving on the separability of all 'flat' hand shapes. This indicates that there might be some merit in utilizing information from HLE. Finally, it can be clearly seen that PGM-HLE produces superior results in terms of separability and cluster interpretability in addition to lowest KL score.

3.4. Spectral clustering on TS-PGM

To investigate contributions of different tensor modes on clustering accuracy, we use spectral clustering (SC), a simple and fast algorithm. Both **CMB** and **UT** datasets contain labels, which we use to evaluate the accuracy as defined as in [10]. In short, cluster class is determined by the labels of majority members, and accuracy is defined as number of correctly clustered data points divided by number of total samples. Results are presented in Tab. 1.

Clustering performance differs across tensor modes. On the entire **CMB** dataset, modes 1 and 3 perform similarly at *75.55%* and *75.44%*, with mode 2 performing worse at *69.77%*. Performances vary in subsets of **CMB**, most likely due to different illumination settings affecting spatio-temporal features. All three modes significantly outperform the baseline at *18.80%*, indicating valuable information contained within them. Furthermore, HLE performs the best compared to individual modes on **CMB**, offering noticeable improvement. In **UT** dataset mode 1 outperforms other two modes and the baseline at *80.90%*. Unlike **CMB**, the nature of modes 1 and 2 is harder to interpret due to the structure of skeletal data. However, it is noticeable that HLE significantly improves the performance of temporal mode from *65.32%* to *77.38%*.

It worth noting that M1 provides superior accuracy compared to HLE when PGM and PGM-HLE are not available. This behavior may happen when some of the classes present very similar shapes where the ordering of the observations over time does not define their semantics.

Unifying information from different tensor modes consistently improves accuracy in all cases, shown on PGM and PGM-HLE performances. In PGM [26], a tensor is represented as a point on product Grassmann manifold, and serves as a baseline for the idea of unifying tensor modes. PGM-HLE offers additional context by utilizing specialized encoding of temporal information, and the improvement is consistent across all datasets.

4. CONCLUSIONS

In this paper we introduced a method for representing temporal tensors based on established multilinear algebra. We use the PGM geometry to naturally unify representations of tensor modes and Hankel-like embedding of temporal information and apply the geodesic distance to investigate the relationship between temporal tensors. We further demonstrate the use of geodesic distance as an general interface for solving optimization and clustering problems.

Using this interface, we performed t-SNE visualizations and spectral clustering of temporal tensor datasets containing video and skeletal data, giving some weight to the strategy of unified representation on PGM, special treatment of temporal information via Hankel-like embeddings and finally the idea of geodesic distance as a general interface for solving various problems. Specifically in the context of video datasets, this approach may prove valuable as it allows simple and fast analysis of data in its raw form, without the need for significant data pre-processing or pre-training of heavy representational models.

As potential future research steps, we will consider several directions. First would be to evaluate the proposed method on various types of multilinear data, such a relational and signal data. Specifically, we believe that unified representation on PGM would be effective in utilizing side information in addition video data, such as sound information, movement information via gyroscope signals, etc. Secondly, we plan to investigate different Riemannian manifolds in order to leverage their different characteristics and similarity metrics that they provide, as potential improvements to the representational aspect of our method. Lastly, a potential future direction includes extending the proposed method to consider applying kernel trick to handle potential non-linearity in tensor modes and other data.

5. ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Numbers JP19H04129, JP19K20335, JP22K17960; and the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) scholarship.

References

[1] Sherif Azary and Andreas Savakis. Grassmannian sparse representations and motion depth surfaces for 3d action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 492–499, 2013. 4

[2] Åke Björck and Gene H Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of computation*, 27(123):579–594, 1973. 2

[3] Salah Bourennane, Caroline Fossati, and Alexis Cailly. Improvement of classification for hyperspectral images based on tensor modeling. *IEEE Geoscience and Remote Sensing Letters*, 7(4):801–805, 2010. 1

[4] Huiyuan Chen and Jing Li. Modeling relational drug-target-disease interactions via tensor factorization with multiple web sources. In *The World Wide Web Conference*, pages 218–227, 2019. 1

[5] Wenwen Ding, Kai Liu, Evgeny Belyaev, and Fei Cheng. Tensor-based linear dynamical systems for action recognition from 3d skeletons. *Pattern Recognition*, 77:75–86, 2018. 1

[6] Kazuhiro Fukui and Atsuto Maki. Difference subspace and its generalization for subspace-based methods. *IEEE transactions on pattern analysis and machine intelligence*, 37(11):2164–2177, 2015. 1, 4

[7] Xizhan Gao, Quansen Sun, Haitao Xu, and Jianqiang Gao. Sparse and collaborative representation based kernel pairwise linear regression for image set classification. *Expert Systems with Applications*, 140:112886, 2020. 2

[8] Bernardo B Gatto, Anna Bogdanova, Lincon S Souza, and Eulanda M dos Santos. Hankel subspace method for efficient gesture representation. In *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2017. 2

[9] Bernardo B Gatto, Eulanda M dos Santos, Alessandro L Koerich, Kazuhiro Fukui, and Waldir SS Junior. Tensor analysis with n-mode generalized difference subspace. *Expert Systems with Applications*, 171:114559, 2021. 2

[10] Bernardo B Gatto, Eulanda M dos Santos, Marco AF Molinetti, and Kazuhiro Fukui. Multilinear clustering via tensor fukunaga–koontz transform with fisher eigenspectrum regularization. *Applied Soft Computing*, page 107899, 2021. 7

[11] Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU press, 2013. 2

[12] Jihun Hamm. *Subspace-based learning with Grassmann kernels*. PhD thesis, University of Pennsylvania, 2008. 2

[13] Mehrtash Harandi, Richard Hartley, Chunhua Shen, Brian Lovell, and Conrad Sanderson. Extrinsic methods for coding and dictionary learning on grassmann manifolds. *International Journal of Computer Vision*, 114(2):113–136, 2015. 2

[14] Mehrtash Harandi, Conrad Sanderson, Chunhua Shen, and Brian C Lovell. Dictionary learning and sparse coding on grassmann manifolds: An extrinsic solution. In *Proceedings of the IEEE international conference on computer vision*, pages 3120–3127, 2013. 2

[15] Chengcheng Jia and Yun Fu. Low-rank tensor subspace learning for rgb-d action recognition. *IEEE Transactions on Image Processing*, 25(10):4641–4652, 2016. 1

[16] Heidi Johansen-Berg and Timothy EJ Behrens. *Diffusion MRI: from quantitative measurement to in vivo neuroanatomy*. Academic Press, 2013. 1

[17] Mohamad Jouni, Mauro Dalla Mura, and Pierre Comon. Hyperspectral image classification using tensor cp decomposition. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 1164–1167. IEEE, 2019. 1

[18] Mohamad Jouni, Mauro Dalla Mura, and Pierre Comon. Hyperspectral image classification based on mathematical morphology and tensor decomposition. *Mathematical Morphology-Theory and Applications*, 4(1):1–30, 2020. 1

[19] Charilaos I Kanatsoulis, Xiao Fu, Nicholas D Sidiropoulos, and Wing-Kin Ma. Hyperspectral super-resolution: A coupled tensor factorization approach. *IEEE Transactions on Signal Processing*, 66(24):6503–6517, 2018. 1

[20] Tae-Kyun Kim, Shu-Fai Wong, and Roberto Cipolla. Tensor canonical correlation analysis for action classification. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 5

[21] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009. 1, 3

[22] Piotr Koniusz, Lei Wang, and Anoop Cherian. Tensor representations for action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1

[23] Damien Letexier, Salah Bourennane, and Jacques Blanc-Talon. Nonorthogonal tensor matricization for hyperspectral image filtering. *IEEE Geoscience and Remote Sensing Letters*, 5(1):3–7, 2008. 1

[24] Binlong Li, Mustafa Ayazoglu, Teresa Mao, Octavia I Camps, and Mario Sznaier. Activity recognition using dynamic subspace angles. In *CVPR 2011*, pages 3193–3200. IEEE, 2011. 2

[25] Yui Man Lui. Advances in matrix manifolds for computer vision. *Image and Vision Computing*, 30(6-7):380–388, 2012. 2

[26] Yui Man Lui. Human gesture recognition on product manifolds. *The Journal of Machine Learning Research*, 13(1):3297–3321, 2012. 1, 7

[27] Yuan Luo, Fei Wang, and Peter Szolovits. Tensor factorization toward precision medicine. *Briefings in bioinformatics*, 18(3):511–514, 2017. 1

[28] Anant Madabhushi and George Lee. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical image analysis*, 33:170–175, 2016. 1

[29] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002. 5

[30] Erkki Oja. *Subspace methods of pattern recognition*, volume 6. John Wiley & Sons, 1983. 1

[31] Thomas Papastergiou, Evangelia I Zacharaki, and Vasileios Megalooikonomou. Tensor decomposition for multiple-instance classification of high-order medical data. *Complexity*, 2018, 2018. 1

[32] Thomas Schultz and Hans-Peter Seidel. Estimating crossing fibers: A tensor decomposition approach. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1635–1642, 2008. 1

[33] Qiquan Shi, Jiaming Yin, Jiajun Cai, Andrzej Cichocki, Tatsuya Yokota, Lei Chen, Mingxuan Yuan, and Jia Zeng. Block hankel tensor arima for multiple short time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5758–5766, 2020. 2

[34] Lincon S Souza, Bernardo B Gatto, and Kazuhiro Fukui. Grassmann singular spectrum analysis for bioacoustics classification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 256–260. IEEE, 2018. 2

[35] Lincon S Souza, Bernardo B Gatto, Jing-Hao Xue, and Kazuhiro Fukui. Enhanced grassmann discriminant analysis with randomized time warping for motion recognition. *Pattern Recognition*, 97:107028, 2020. 2

[36] Lincon S Souza, Naoya Sogi, Bernardo B Gatto, Takumi Kobayashi, and Kazuhiro Fukui. An interface between grassmann manifolds and vector spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 846–847, 2020. 2

[37] Lincon S Souza, Naoya Sogi, Bernardo B Gatto, Takumi Kobayashi, and Kazuhiro Fukui. Grassmannian learning mutual subspace method for image set recognition. *arXiv preprint arXiv:2111.04352*, 2021. 2

[38] GW Stewart and JG Sun. Computer science and scientific computing. matrix perturbation theory, 1990. 2

[39] Chendra Hadi Suryanto, Jing-Hao Xue, and Kazuhiro Fukui. Randomized time warping for motion recognition. *Image and Vision Computing*, 54:1–11, 2016. 2

[40] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 5

[41] Claudio Varini, Andreas Degenhard, and Tim W Nattkemper. Isolle: Lle with geodesic distance. *Neurocomputing*, 69(13-15):1768–1771, 2006. 2

[42] Long Wang, JL Wang, ZL Cheng, L Ran, and Z Yin. Personalized medicine recommendation based on tensor decomposition. *Comput Sci*, 42:225–229, 2015. 1

[43] Satosi Watanabe and Nikhil Pakvasa. Subspace method of pattern recognition. In *Proc. 1st. IJCPR*, pages 25–32, 1973. 1

[44] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. *Distill*, 1(10):e2, 2016. 7

[45] Lu Xia, Chia-Chih Chen, and Jake K Aggarwal. View invariant human action recognition using histograms of 3d joints. In *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 20–27. IEEE, 2012. 5

[46] Yeyang Yu, Jin Jin, Feng Liu, and Stuart Crozier. Multidimensional compressed sensing mri using tensor decomposition-based sparsifying transform. *PloS one*, 9(6):e98441, 2014. 1