Tensor Discriminant Analysis on Grassmann Manifold with Application to Video based Human Action Recognition

Cagri Ozdemir[a], Randy C. Hoover[a], Kyle Caudle[b], Karen Braman[b]

[a]*Department of Computer Science and Engineering, South Dakota Mines, 501 E St Joseph St, Rapid City, SD 57701, USA*
[b]*Department of Mathematics, South Dakota Mines, 501 E St Joseph St, Rapid City, SD 57701, USA*

Abstract

Representing videos as linear subspaces on Grassmann manifolds has made great strides in action recognition problems. Recent studies have explored the convenience of discriminant analysis by making use of Grassmann kernels. However, traditional methods rely on the matrix representation of videos based on the temporal dimension and suffer from not considering the two spatial dimensions. To overcome this problem, we keep the natural form of videos by representing video inputs as multidimensional arrays known as tensors and propose a tensor discriminant analysis approach on Grassmannian manifolds. Because matrix algebra does not handle tensor data, we introduce a new Grassmann projection kernel based on the tensor-tensor decomposition and product. Experiments with human action databases show that the proposed method performs well compared with the state-of-the-art algorithms.

Keywords: Grassmann Discriminant Analysis, Action recognition, Tensor singular value decomposition, Tensor eigendecomposition.

1. Introduction

Human action recognition has been widely used in many application areas, such as intelligent video surveillance [1], human-computer interfaces [2], and identity recognition [3]. In the literature, many human action recognition algorithms have been built upon modeling video sequences as linear subspaces [4, 5, 6, 7, 8, 9, 10]. As human action videos are 3-D data objects capturing both spatial and temporal information from human actions, the high-order singular value decomposition (HOSVD) provides a way that each mode can be analyzed separately. For example, as outline in tangent bundle architecture (TB) [7], data tensors were factorized using the HOSVD. Then each factor was projected onto a tangent space and the intrinsic distance was computed from a tangent bundle for action classification. Moreover, product Grassmann manifold (PGM) [8] and n-mode generalized difference subspaces (n-mode GDS) [10] methods consider the geodesic distances on the Grassmann manifold in a different perspective, where the HOSVD was used to capture information from both spatial and temporal information from human actions.

As the subspaces of a Euclidean space lie on a special type of Riemannian manifolds, the Grassmann manifold, which has a nonlinear structure, cannot be analyzed using Euclidean geometry. To solve this problem, in [4], the Projection kernel and the Binet-Cauchy kernel have been introduced to embed the Grassmann manifold into a Hilbert space which has a Euclidean geometry. Then Grassmann discriminant analysis (GDA) was proposed by using kernel LDA with the Grassmann kernels. Similarly, graph-embedded Grassmann discriminant analysis (GGDA) method [5] was proposed based on a graph-embedding framework. Furthermore, an extended family of Grassmann kernels has been introduced in [11]. Despite their success, both GDA and GGDA rely on the matrix representation of videos based on the temporal dimension and suffer from not considering the two spatial dimensions. Additionally, as each video frame needs to be "vectorized" into a column vector, the natural representation of the sample eliminates the spatial correlation within each sample and leads to the estimation of the large number of parameters.

*Corresponding author email: cagri.ozdemir@mines.sdsmt.edu
Email addresses: cagri.ozdemir@mines.sdsmt.edu (Cagri Ozdemir), randy.hoover@sdsmt.edu (Randy C. Hoover), kyle.caudle@sdsmt.edu (Kyle Caudle), karen.braman@sdsmt.edu (Karen Braman)

Due to these issues, particularly when dealing with higher-order data, there has been a growing interest in multilinear subspace learning (MSL) that keeps the natural representation of the multidimensional arrays (commonly referred to as tensors). Tensors provide a natural framework for representing higher-order data and the tensor singular value decomposition (**t-SVD**) [12, 13, 14] has been widely used as multilinear extension of linear algebra tools. Fundamental to **t-SVD** is the defined multiplication operator on third-order tensors (t-product) based upon Fourier theory and an algebra of circulants [15, 16]. As the t-product is a convolution-like operation that can be implemented using the Fast Fourier Transform (FFT), variations on the classical t-product have been investigated in [17] where it is shown that a family tensor-tensor products can be defined directly in a transform domain for an arbitrary invertible linear transform. Most recently, fast algorithms have been developed for the t-product and t-SVD [18, 19]. Furthermore, the t-product (and the many variations therein) have been applied to tensor completion [20, 21] and image processing [22, 23].

In this paper, we first introduce the discrete wavelet transform (DWT) based third-order tensor definitions and mathematical operations using a special structured block matrix that are computationally more efficient compared to the FFT and the discrete cosine transform (DCT) based third-order operators [24, 25, 17]. To capture the information from the two spatial dimensions and the temporal dimension of human action videos, we propose the DWT-based 3-mode tensor singular value decomposition (3-mode $\mathbf{t\text{-}SVD_w}$). In doing so, we obtain multilinear subspaces representing each mode. We further consider subspaces of each mode separately and generate feature representation in the Euclidean space. Since matrix algebra does not handle tensor data, we introduce a new well-defined positive definite Grassmann tensor projection kernel, which is built upon the tensor-tensor decomposition and product, to embed the Grassmann manifold to a Hilbert space. By doing this, we construct independent vector spaces corresponding each mode. Finally, we fuse these vector spaces by applying kernel-based multilinear discriminant analysis (KMLDA). As a result, we demonstrate the proposed framework called tensor Grassmann discriminant analysis (TGDA) by using Grassmann tensor projection kernel and by applying KMLDA to human action classification problems.

The contributions of the proposed work can be summarized as follows:

- Introduction of an improved version of **t-SVD** called 3-mode $\mathbf{t\text{-}SVD_w}$.
- A novel tensor-based Grassmann kernel function and necessary kernel validity conditions are proved.
- A novel formulation for kernel-based multilinear discriminant analysis.
- Competitive classification achievement on human action recognition.

The rest of the paper is organized as follows. In Section 2, we discuss the mathematical foundations of the tensor operators and the tensor-tensor decomposition. In Section 3, we introduce the proposed method. Section 4 reports our experimental results, and Section 5 concludes the paper.

2. Mathematical Background

In this section, we define the wavelet transform-based third-order tensor definitions and mathematical operations using a specially structured block matrix. Fundamental to the results presented in this section is motivated by the tensor definitions based upon the Fourier theory and the algebra of circulant as outlined in [25, 24, 12, 14, 13, 26, 16, 27, 19].

2.1. Tensor operators

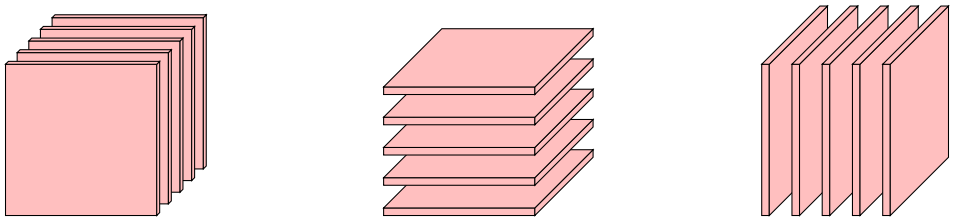
In the wavelet transform, a signal in the time domain is decomposed by passing it through high-pass filter (resulting in detail coefficients) and low-pass filter (resulting in approximation coefficients) to produce low-pass and high-pass wavelet coefficients referred to as a "*level-1 decomposition*". The low-pass version can be further decomposed by again passing it to a set of low-pass and high pass filters referred to as "*level-2 decomposition*". This process can be further continued to a pre-defined level as outlined in [28, 29]. While there are many different types of wavelets, arguably the most common are the Haar wavelet [30] and the Daubechies wavelet [31]. In our work, we use the Haar wavelet due to its low computation cost and simplicity to apply as compared to other wavelets. The discrete Haar

wavelet transform can be expressed in matrix form for each level decomposition [32]. Thus, the discrete Haar wavelet forward transformation matrix for level-1 decomposition can be written as:

$$H = \begin{bmatrix} h_0 & h_1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & h_0 & h_1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \cdots & \cdots & \cdots & h_0 & h_1 \\ g_0 & g_1 & 0 & \cdots & \cdots & \cdots & 0 & 0 \\ 0 & 0 & g_0 & g_1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \cdots & \cdots & \cdots & g_0 & g_1 \end{bmatrix}, \tag{1}$$

where h_0 and h_1 are scaling function coefficients, whereas g_0 and g_1 are wavelet function coefficients.

It will be convenient to break a tensor $\mathcal{A}$ in $\mathbb{R}^{\ell \times m \times n}$ up into various slices and to have an indexing on those. The i^{th} frontal slice will be denoted $\mathcal{A}^{(i)}$, the j^{th} horizontal slice will be denoted $\mathcal{A}_{(j)}$, and the k^{th} lateral slice will be denoted $\vec{\mathcal{A}}_{(k)}$. In terms of MATLAB indexing notation, this means $\mathcal{A}^{(i)} \equiv \mathcal{A}(:,:,i)$, $\mathcal{A}_{(j)} \equiv \mathcal{A}(j,:,:)$, and $\vec{\mathcal{A}}_{(k)} \equiv \mathcal{A}(:,k,:)$.



(a) (b) (c)

Figure 1: (a) Frontal, (b) horizontal, and (c) lateral slices of a third-order tensor.

In order to discuss our new definitions we must first introduce the block matrix that can be diagonalized by the discrete Haar wavelet transform matrix H (illustrated in (1)). We call this block matrix the "*block dwt matrix*", denoted by **bdwt** for short. For example, consider the tensor $\mathcal{A} \in \mathbb{R}^{\ell \times m \times n}$ with $\ell \times m$ frontal slices then $\mathbf{bdwt}(\mathcal{A})$ can be written as follows:

$$\mathbf{bdwt}(\mathcal{A}) = \begin{bmatrix} \mathcal{A}^{(1)} & \mathcal{A}^{(2)} & 0 & 0 & \cdots & 0 \\ \mathcal{A}^{(2)} & \mathcal{A}^{(1)} & 0 & 0 & \cdots & \vdots \\ 0 & 0 & \mathcal{A}^{(3)} & \mathcal{A}^{(4)} & \cdots & \vdots \\ 0 & 0 & \mathcal{A}^{(4)} & \mathcal{A}^{(3)} & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \mathcal{A}^{(n-1)} & \mathcal{A}^{(n)} \\ 0 & 0 & \cdots & \cdots & \mathcal{A}^{(n)} & \mathcal{A}^{(n-1)} \end{bmatrix}. \tag{2}$$

A new block-diagonal form (3) can be constructed via left and right multiplication by a DWT matrix (1).

$$(H_n \otimes I_\ell) \cdot \mathbf{bdwt}(\mathcal{A}) \cdot (H_n^T \otimes I_m) = \begin{bmatrix} D_1 & & & \\ & D_2 & & \\ & & \ddots & \\ & & & D_n \end{bmatrix}, \tag{3}$$

where each of the D_i is a $\ell \times m$ matrix, I_ℓ is a $\ell \times \ell$ identity matrix, I_m is a $m \times m$ identity matrix, H_n is the $n \times n$ DWT matrix defined in (1), H_n^T is its transpose, and $\otimes$ is the Kronecker product.

Definition 1. An element $\mathbf{c} \in \mathbb{R}^{1 \times 1 \times n}$ is called a **tubal-scalar** of length n.

Definition 2. If $\mathcal{A} \in \mathbb{R}^{\ell \times m \times n}$, then **unfold**($\mathcal{A}$) takes tensor $\mathcal{A}$ and returns a block $\ell n \times m$ matrix.

$$\textbf{unfold}(\mathcal{A}) = \begin{bmatrix} \mathcal{A}^{(1)} \\ \mathcal{A}^{(2)} \\ \vdots \\ \mathcal{A}^{(n)} \end{bmatrix}.$$

The operation that takes **unfold**($\mathcal{A}$) back to tensor form is the **fold** operator:

$$\mathcal{A} = \textbf{fold}(\textbf{unfold}(\mathcal{A})).$$

Definition 3. If $\mathcal{A} \in \mathbb{R}^{\ell \times m \times n}$, then the **unbdwt** operator takes matrix **bdwt**($\mathcal{A}$) and returns tensor $\mathcal{A}$.

$$\mathcal{A} = \textbf{unbdwt}(\textbf{bdwt}(\mathcal{A})).$$

Definition 4. Let $\mathcal{A} \in \mathbb{R}^{\ell \times m \times n}$ and $\mathcal{B} \in \mathbb{R}^{m \times \ell \times n}$. Then the wavelet product denoted by $\mathcal{A} *_w \mathcal{B} \in \mathbb{R}^{\ell \times \ell \times n}$ is defined as:

$$\mathcal{A} *_w \mathcal{B} = \textbf{fold}(\textbf{bdwt}(\mathcal{A}) \cdot \textbf{unfold}(\mathcal{B})),$$

where "$\cdot$" is the standard matrix multiplication.

Definition 5. Let $\mathcal{A} \in \mathbb{R}^{\ell \times m \times n}$. Then the tensor transpose of the tensor $\mathcal{A}$ denoted by $\mathcal{A}^T \in \mathbb{R}^{m \times \ell \times n}$ is defined as:

$$\mathcal{A}^T = \textbf{unbdwt}\Big((\textbf{bdwt}(\mathcal{A}))^T\Big).$$

Definition 6. Let $\mathcal{A} \in \mathbb{R}^{m \times m \times n}$. Then the tensor inverse of the tensor $\mathcal{A}$ denoted by $\mathcal{A}^{-1} \in \mathbb{R}^{m \times m \times n}$ is defined as:

$$\mathcal{A}^{-1} = \textbf{unbdwt}\Big((\textbf{bdwt}(\mathcal{A}))^{-1}\Big).$$

Definition 7. The identity tensor $\mathcal{I} \in \mathbb{R}^{m \times m \times n}$ is the tensor whose frontal slice is the $m \times m$ identity matrix in the transform domain,

$$\mathcal{I} = \tilde{\mathcal{I}} \times_3 H_n^{-1},$$

where $\tilde{\mathcal{I}}(:,:,i) = I$ for $i = 1, \ldots, n$ and I is the $m \times m$ identity matrix. H_n is the $n \times n$ Haar wavelet level-1 transformation matrix defined in (1). We note that $\times_3$ is a mode-3 product [33, 17]. This operation has the same effect as taking the inverse wavelet transform along each tubal-scalar of $\tilde{\mathcal{I}}$.

Definition 8. The tensor norm used through this paper is the Frobenious norm which for the tensor $\mathcal{A} \in \mathbb{R}^{\ell \times m \times n}$ is given by:

$$\|\mathcal{A}\|_F = \sqrt{\sum_{i=1}^{\ell} \sum_{j=1}^{m} \sum_{k=1}^{n} (\mathcal{A}(i,j,k))^2}.$$

Definition 9. An idempotent tensor $\mathcal{A} \in \mathbb{R}^{m \times m \times n}$ is the tensor which, when multiplied by itself, yields itself.

$$\mathcal{A} = \mathcal{A} *_w \mathcal{A}.$$

2.2. Wavelet Tensor Eigendecomposition and Singular Value Decomposition

Motivated by the tensor eigendecomposition based on the Fourier transform-based tensor operators in [34], we will introduce the wavelet tensor eigendecomposition referred to as the **t-eig$_\mathbf{w}$** and the wavelet tensor eigendecomposition referred to as the **t-SVD$_\mathbf{w}$** using the special block structure and tensor operators given in Section 2.1.

Theorem 1. Let $\mathcal{A} \in \mathbb{R}^{m \times m \times n}$, then the **t-eig$_\mathbf{w}$** of the tensor $\mathcal{A}$ can be factored as:

$$\mathcal{A} = \mathcal{P} *_w \mathcal{D} *_w \mathcal{P}^{-1},$$

where $\mathcal{P} \in \mathbb{R}^{m\times m\times n}$ is a non-singular and $\mathcal{D} \in \mathbb{R}^{m\times m\times n}$ is a f-diagonal tensor such that the frontal slices are diagonal. A graphical illustration of the **t-eig$_\mathbf{w}$** is shown in Fig.

Proof. Recall equation (3):

$$(H_n \otimes I_m) \cdot \mathbf{bdwt}(\mathcal{A}) \cdot (H_n^T \otimes I_m) = \begin{bmatrix} D_1 & & & \\ & D_2 & & \\ & & \ddots & \\ & & & D_n \end{bmatrix}.$$

To construct the **t-eig$_\mathbf{w}$**, the matrix eigenvalue decomposition is performed on each of the D_i as $D_i = P_i \Sigma_i P_i^{-1}$. Then we can write:

$$\begin{bmatrix} D_1 & & & \\ & \ddots & & \\ & & D_n \end{bmatrix} = \begin{bmatrix} P_1 & & & \\ & \ddots & & \\ & & P_n \end{bmatrix} \begin{bmatrix} \Sigma_1 & & & \\ & \ddots & & \\ & & \Sigma_n \end{bmatrix} \begin{bmatrix} P_1^{-1} & & & \\ & \ddots & & \\ & & P_n^{-1} \end{bmatrix}.$$

Applying $(H_n^T \otimes I_m)$ to the left and $(H_n \otimes I_m)$ to the right of each of the block diagonal matrices on the right hand side results in each being the **bdwt** structure. We can use the **unbdwt** operator given in **Definition 3** to take them back into tensor form as following:

$$\mathcal{P} = \mathbf{unbdwt}\left((H_n^T \otimes I_m) \begin{bmatrix} P_1 & & \\ & \ddots & \\ & & P_n \end{bmatrix} (H_n \otimes I_m)\right),$$

$$\mathcal{D} = \mathbf{unbdwt}\left((H_n^T \otimes I_m) \begin{bmatrix} \Sigma_1 & & \\ & \ddots & \\ & & \Sigma_n \end{bmatrix} (H_n \otimes I_m)\right),$$

$$\mathcal{P}^{-1} = \mathbf{unbdwt}\left((H_n^T \otimes I_m) \begin{bmatrix} P_1^{-1} & & \\ & \ddots & \\ & & P_n^{-1} \end{bmatrix} (H_n \otimes I_m)\right),$$

Therefore, that results in the decomposition:

$$\mathcal{A} = \mathcal{P} *_w \mathcal{D} *_w \mathcal{P}^{-1}.$$

□

Theorem 2. Let $\mathcal{A} \in \mathbb{R}^{\ell\times m\times n}$, then the **t-SVD$_\mathbf{w}$** of the tensor $\mathcal{A}$ can be factored as:

$$\mathcal{A} = \mathcal{U} *_w \mathcal{S} *_w \mathcal{V}^T,$$

where $\mathcal{U} \in \mathbb{R}^{\ell\times\ell\times n}$ is a left-orthogonal tensor, $\mathcal{V} \in \mathbb{R}^{m\times m\times n}$ is a right-orthogonal tensor and $\mathcal{S} \in \mathbb{R}^{m\times m\times n}$ is a f-diagonal tensor such that the frontal slices are diagonal. A graphical illustration of the **t-SVD$_\mathbf{w}$** is shown in Fig.

Proof. Recall equation (3):

$$(H_n \otimes I_\ell) \cdot \mathbf{bdwt}(\mathcal{A}) \cdot (H_n^T \otimes I_m) = \begin{bmatrix} D_1 & & & \\ & D_2 & & \\ & & \ddots & \\ & & & D_n \end{bmatrix}.$$

To construct the **t-SVD$_{\mathbf{w}}$**, the matrix eigenvalue decomposition is performed on each of the D_i as $D_i = U_i \Sigma_i V_i^T$. Then we can write:

$$\begin{bmatrix} D_1 & & \\ & \ddots & \\ & & D_n \end{bmatrix} = \begin{bmatrix} P_1 & & \\ & \ddots & \\ & & P_n \end{bmatrix} \begin{bmatrix} \Sigma_1 & & \\ & \ddots & \\ & & \Sigma_n \end{bmatrix} \begin{bmatrix} V_1^T & & \\ & \ddots & \\ & & V_n^T \end{bmatrix}.$$

Similarly to the proof of **Theorem 1**, applying $(H_n^T \otimes I_\ell)$ to the left and $(H_n \otimes I_m)$ to the right of each of the block diagonal matrices on the right hand side results in each being the **bdwt** structure. We can use the **unbdwt** operator given in **Definition 3** to take them back into tensor form. Showing $\mathcal{U} *_w \mathcal{U}^T = \mathcal{U}^T *_w \mathcal{U} = \mathcal{I}$ and $\mathcal{V} *_w \mathcal{V}^T = \mathcal{V}^T *_w \mathcal{V} = \mathcal{I}$ completes the proof. □

For computational efficiency, we can compute both the tensor eigendecomposition and the tensor singular value decomposition for any invertible transforms using spectral domain operations similar to the computation of the t-SVD using the FFT in place of spatial domain operations [12, 14, 24, 17, 35]. Previously, the FFT-based tensor eigendecomposition and tensor singular value decomposition have been introduced in [12, 17, 34]. However, the DWT has time complexity of $O(N)$, whereas the FFT and DCT are both $O(NlogN)$, hence the DWT provides significant reduction in computational complexity. TABLE 1 shows the time complexity of the tensor eigendecomposition and the tensor singular value decomposition of tensor $\mathcal{A} \in \mathbb{R}^{m \times m \times n}$ based on three most widely used invertable linear transforms, it is clearly seen that both the DWT-based tensor eigendecomposition (**t-eig$_{\mathbf{w}}$**) and the DWT-based tensor singular value decomposition (**t-SVD$_{\mathbf{w}}$**) are the fastest method among others.

Table 1: The time complexity of the FFT, DCT, and DWT based tensor eigendecomposition and tensor singular value decomposition.

Tensor $\mathcal{A} \in \mathbb{R}^{m \times m \times n}$					
FFT	$O(m^2 nlog(n))$	DCT	$O(m^2 nlog(n))$	DWT	$O(m^2 n)$
t-eig after FFT	$O(2m^3)$	**t-eig** after DCT	$O(m^3)$	**t-eig** after DWT	$O(m^3)$
t-eig with FFT	$O(m^2 nlog(n)) + O(2m^3)$	**t-eig** with DCT	$O(m^2 nlog(n)) + O(m^3)$	**t-eig** with DWT	$O(m^2 n) + O(m^3)$
t-SVD after FFT	$O(2nm^3)$	**t-SVD** after DCT	$O(nm^3)$	**t-SVD** after DWT	$O(nm^3)$
t-SVD with FFT	$O(m^2 nlog(n)) + O(2nm^3)$	**t-SVD** with DCT	$O(m^2 nlog(n)) + O(nm^3)$	**t-SVD** with DWT	$O(m^2 n) + O(nm^3)$

3. Proposed Method

In this section, we first introduce tensor subspaces as elements of a Grassmann manifold. We then introduce a tensor-based Grassmann kernel function to embed the Grassmann manifold into a Hilbert space. Finally, we turn our attention to our proposed model for formulating a kernel-based multilinear discriminant analysis in a vector space.

3.1. Representing Subspaces on the Product Manifold

A Grassmann manifold $\mathcal{G}(d, m)$ is defined as the set of m dimensional linear subspaces of $\mathbb{R}^d$ [36][1]. Given a linear subspace $X \in \mathbb{R}^{d \times m}$, which contains an orthonormal set, can be represented as a point on the Grassmann manifold $\mathcal{G}(d, m)$. The collection of all possible permutations of $X \in \mathbb{R}^{d \times m}$ forms a manifold structure defined by $\mathcal{G}(d, m)$.

As outlined in Section 2.2, to construct the **t-SVD$_{\mathbf{w}}$**, the matrix singular value is performed on each element of the block diagonal form. A third-order tensor $\mathcal{A} \in \mathbb{R}^{\ell \times m \times n}$ can be diagonalized as:

$$\mathcal{A} = \mathcal{U} *_w \mathcal{S} *_w \mathcal{V}^T$$

[1]Note that while we generally use upper-case calligraphic letters to denote tensors, to keep consistent with the literature, we will denote a Grassmann manifold using an upper case calligraphic $\mathcal{G}$.

where $\mathcal{U} \in \mathbb{R}^{\ell \times \ell \times n}$ and $\mathcal{V} \in \mathbb{R}^{m \times m \times n}$ are orthogonal tensors. Each element of an orthogonal tensor is a an orthogonal matrix in so-called transform domain. Transform domain representation of a tensor can be obtained via left and right multiplication by a DWT matrix given in (3). Hence, we can write:

$$(H_n \otimes I_\ell) \cdot \mathbf{bdwt}(\mathcal{U}) \cdot (H_n^T \otimes I_\ell) = \begin{bmatrix} U_1 & & & \\ & U_2 & & \\ & & \ddots & \\ & & & U_n \end{bmatrix},$$

$$(H_n \otimes I_m) \cdot \mathbf{bdwt}(\mathcal{V}) \cdot (H_n^T \otimes I_m) = \begin{bmatrix} V_1 & & & \\ & V_2 & & \\ & & \ddots & \\ & & & V_n \end{bmatrix}.$$

where $U_1, U_2, \ldots, U_n$ are $\ell \times \ell$ orthogonal matrices ($U_1, U_2, \ldots, U_n \in O(\ell)$), whereas $V_1, V_2, \ldots, V_n$ are $m \times m$ orthogonal matrices ($V_1, V_2, \ldots, V_n \in O(m)$). Similarly, k dimensional subspaces of these orthogonal tensors ($\bar{\mathcal{U}} \in \mathbb{R}^{\ell \times k \times n}$ and $\bar{\mathcal{V}} \in \mathbb{R}^{m \times k \times n}$) provide $\ell \times k$ and $m \times k$ matrices in the transform domain respectively, where each matrix consists of a k dimensional orthonormal set. These matrices are elements of $\mathcal{G}(\ell, k)$, and $\mathcal{G}(m, k)$. Therefore, an orthogonal tensor or k dimensional subspace of an orthogonal tensor can be considered as a bundle of the elements of a Grassmann manifold. It is important to note that all elements of an orthogonal tensor live on the same manifold.

Action videos are multidimensional data and can be naturally represented as third-order data tensors that generally have two spatial modes and a temporal one. Since action videos provide discriminative information in each mode, each mode must be analyzed independently [6, 8]. However, one of the fundamental drawbacks associated with computing the **t-SVD$_\mathbf{w}$** in this fashion is the choice of flattening the data tensor through the **bdwt**$(\cdot)$ operator. By construction, **bdwt**$(\cdot)$ operates on the frontal slices of a third-order tensor (mode-3), however this "choice" is somewhat arbitrary (e.g. we could just as easily reformulate the problem to operate on horizontal slices (mode-1) or lateral slices (mode-2)). As such, while we capture correlations in the video data along a specific mode, we neglect the correlations along other two modes. As shown in Figure 1, there are three ways to slice a third-order tensor. In order to capture distinct properties from tensor data, we employ the 3-mode **t-SVD$_\mathbf{w}$** that produces three different sets of subspaces by operating on frontal, horizontal, and lateral slices independently. In order to stick with the tensor operators defined in Section 2.1, we define **swapmodes** $(\cdot)$ in **Definition 10**, which interchanges two modes of a third-order tensor.

Definition 10. Let $\mathcal{A} \in \mathbb{R}^{\ell \times m \times n}$. Then **swapmodes**$(\mathcal{A}, a, b)$ rearranges tensor $\mathcal{A}$ by interchanging given mode-a and mode-b.

$$\mathbf{swapmodes}(\mathcal{A}, 1, 3) = \mathcal{B} \in \mathbb{R}^{n \times m \times \ell},$$
$$\mathbf{swapmodes}(\mathcal{A}, 2, 3) = C \in \mathbb{R}^{\ell \times n \times m}.$$

As such, we can treat the horizontal and the lateral slices of a third-order tensor as frontal slices using **swapmodes** $(\cdot)$ operator. Given a third-order tensor $\mathcal{A}$, the 3-mode **t-SVD$_\mathbf{w}$** decompose tensor $\mathcal{A}$ as follows:

$$\mathbf{t\text{-}SVD_w}(\mathcal{A}) = {}^1\mathcal{U} *_w {}^1\mathcal{S} *_w {}^1\mathcal{V}^T, \tag{4}$$
$$\mathbf{t\text{-}SVD_w}(\mathbf{swapmodes}(\mathcal{A}, 1, 3)) = {}^2\mathcal{U} *_w {}^2\mathcal{S} *_w {}^2\mathcal{V}^T, \tag{5}$$
$$\mathbf{t\text{-}SVD_w}(\mathbf{swapmodes}(\mathcal{A}, 2, 3)) = {}^3\mathcal{U} *_w {}^3\mathcal{S} *_w {}^3\mathcal{V}^T, \tag{6}$$

where the **t-SVD$_\mathbf{w}$** is computed by operating on the frontal slices, the horizontal slices, and the lateral slices of tensor $\mathcal{A}$ respectively. We use 1, 2, and 3 left superscript to denote products of mode-1 **t-SVD$_\mathbf{w}$**, mode-2 **t-SVD$_\mathbf{w}$**, and mode-3 **t-SVD$_\mathbf{w}$** respectively.

In action recognition studies, the Tucker decomposition is operated on the unfolded modes via matrix unfolding in which the variation of each mode is captured by the Tucker decomposition [6, 8, 10]. However, for large tensors, "unfolding" to compute the n-mode product at each mode results in fat matrices dominated by a single dimension. We can then assume that the dimension of the columns is greater than the dimension of the rows due to the nature

of matrix unfolding for action videos. This implies that the unfolded matrix only spans the row dimension. Thus, subspaces that only span row space are used to represent data tensors on a Grassmann manifold. In our approach, we benefit from representing not only the spaces span row spaces (${}^1\mathcal{V}, {}^2\mathcal{V}, {}^3\mathcal{V}$), but also the subspaces span column spaces (${}^1\mathcal{U}, {}^2\mathcal{U}, {}^3\mathcal{U}$) on Grassmann manifolds.

3.2. Grassmann Tensor Discriminant Analysis

As mentioned earlier, a consistent approach to perform discriminant analysis for linear subspaces is to embed the Grassmann manifold into a Hilbert space using a Grassmann kernel. While Grassmann kernel functions have been recently introduced [4, 37, 11], their applicability is limited by the fact that they only operates on matrices (second-order tensors). In this subsection, we first introduce a positive definite Grassmann kernel function for third-order subspaces and then formulate a discriminant analysis in an unconventional space provided by this kernel function.

3.2.1. Grassmann Tensor Projection Kernel

Let $\mathcal{U}_1, \mathcal{U}_2 \in \mathbb{R}^{m\times p\times \ell}$ two multilinear subspaces obtained by **t-SVD$_{\mathbf{w}}$**. It is important to note that both subspaces belong to the same mode **t-SVD$_{\mathbf{w}}$**. As illustrated in Equation (3), they can be diagonalized as:

$$(H_\ell \otimes I_m)\cdot \mathbf{bdwt}(\mathcal{U}_1)\cdot (H_\ell^T \otimes I_p) = \begin{bmatrix} U_{1_1} & & & \\ & U_{1_2} & & \\ & & \ddots & \\ & & & U_{1_\ell} \end{bmatrix},$$

$$(H_\ell \otimes I_m)\cdot \mathbf{bdwt}(\mathcal{U}_2)\cdot (H_\ell^T \otimes I_p) = \begin{bmatrix} U_{2_1} & & & \\ & U_{2_2} & & \\ & & \ddots & \\ & & & U_{2_\ell} \end{bmatrix},$$

where each diagonal element is a $m \times p$ matrices. The Frobenius norm is denoted as:

$$\|\mathcal{U}_1^T *_w \mathcal{U}_2\|_F^2 = \sum_{k=1}^{\ell} \|U_{1_k}^T U_{2_k}\|_F^2 = \sum_{k=1}^{\ell} \mathrm{tr}\left(U_{1_k} U_{1_k}^T U_{2_k} U_{2_k}^T\right).$$

As outlined in [4, 37, 11], a function $k : \mathcal{G}(m,p)\times\mathcal{G}(m,p)\to\mathbb{R}$ is a Grassmannian kernel if it is well-defined and positive definite. Therefore:

Proposition 1. The tensor projection kernel

$$k_p(\mathcal{U}_1,\mathcal{U}_2) = \|\mathcal{U}_1^T *_w \mathcal{U}_2\|_F^2$$

is a Grassmann kernel.

Proof. The kernel is well-defined if it satisfies two conditions.

1) Positive definiteness

The positive definiteness follows from the properties of the Frobenius norm. For all $\mathcal{U}_i,\ldots,\mathcal{U}_n$ (the diagonal elements $U_{k_i} \in \mathcal{G}$) and $c_1,\cdots,c_n (c_i \in \mathbb{R})$ for any $n \in \mathbb{N}$, we have

$$\begin{aligned}
&\sum_{ij} c_i c_j \|\mathcal{U}_i^T *_w \mathcal{U}_j\|_F^2 = \sum_{ij} c_i c_j \sum_{k=1}^{\ell} \mathrm{tr}(U_{i_k} U_{i_k}^T U_{j_k} U_{j_k}^T),\\
&= \mathrm{tr}\left(\sum_{ij}\sum_{k=1}^{\ell} c_i c_j U_{i_k} U_{i_k}^T U_{j_k} U_{j_k}^T\right) = \sum_{k=1}^{\ell} \mathrm{tr}\left(\sum_{ij} c_i c_j U_{i_k} U_{i_k}^T U_{j_k} U_{j_k}^T\right),\\
&= \sum_{k=1}^{\ell} \mathrm{tr}\left(\sum_{i} c_i U_{i_k} U_{i_k}^T\right)^2 = \sum_{k=1}^{\ell} \|\sum_{i} c_i U_{i_k} U_{i_k}^T\|_F^2 \geq 0.
\end{aligned}$$

2)Invariant to different representation
$k_p(\mathcal{U}_i, \mathcal{U}_j) = k_p(\mathcal{U}_i *_w \mathcal{Y}_i, \mathcal{U}_j *_w \mathcal{Y}_j)$ for any $\mathcal{Y}_i, \mathcal{Y}_j$ third-order orthogonal tensors.

$$
\begin{aligned}
k_p(\mathcal{U}_i *_w \mathcal{Y}_i, \mathcal{U}_j *_w \mathcal{Y}_j) &= \sum_{k=1}^{\ell} \|Y_{i_k}^T U_{i_k}^T U_{j_k} Y_{j_k}\|_F^2, \\
&= \sum_{k=1}^{\ell} \operatorname{tr}\left(Y_{j_k}^T U_{j_k}^T U_{i_k} Y_{i_k} Y_{i_k}^T U_{i_k}^T U_{j_k} Y_{j_k} \right), \\
&= \sum_{k=1}^{\ell} \operatorname{tr}\left(Y_{j_k} Y_{j_k}^T U_{j_k}^T U_{i_k} Y_{i_k} Y_{i_k}^T U_{i_k}^T U_{j_k} \right), \\
&= \sum_{k=1}^{\ell} \operatorname{tr}\left(U_{j_k}^T U_{i_k} U_{i_k}^T U_{j_k} \right) = \sum_{k=1}^{\ell} \operatorname{tr}\left(U_{i_k} U_{i_k}^T U_{j_k} U_{j_k}^T \right), \\
&= \sum_{k=1}^{\ell} \|U_{i_k}^T U_{j_k}\|_F^2 = \|\mathcal{U}_i^T *_w \mathcal{U}_j\|_F^2 = k_p(\mathcal{U}_i, \mathcal{U}_j).
\end{aligned}
$$

□

3.2.2. Overview of Multilinear Discriminant Analysis

To keep the current work self-contained, we present a brief overview of the work on MLDA outlined in [34]. However, unlike MLDA in [34] (which was based Fourier theory), we use the wavelet transform-based operators and definitions as outlined in Section 2.

Consider the situation where we have a collection of ℓ image samples, each of size $m \times n$ pixels. We can construct our data tensor $\mathcal{A} \in \mathbb{R}^{m \times \ell \times n}$ where each lateral slice of $\mathcal{A}$ (i.e., $\vec{\mathcal{A}}_{(i)}$, for $i = 1, 2, \ldots, \ell$) is an $m \times n$ sample image. From this construction, the within-class scatter tensor can be written as:

$$
\mathcal{S}_w = \sum_{i=1}^{c} \sum_{\vec{\mathcal{A}}_{(j)} \in c_i} (\vec{\mathcal{A}}_{(j)} - \vec{\mathcal{M}}_{(i)}) *_w (\vec{\mathcal{A}}_{(j)} - \vec{\mathcal{M}}_{(i)})^T,
$$

where c is the total number of classes, $\vec{\mathcal{A}}_{(j)} \in \mathbb{R}^{m \times 1 \times n}$ is the j^{th} lateral slice of class i denoted by c_i, and $\vec{\mathcal{M}}_{(i)} \in \mathbb{R}^{m \times 1 \times n}$ is the mean of class c_i. The transpose operator and the multiplication operator are outlined in **Definition 5** and **Definition 4** respectively. We define the between-class scatter tensor as:

$$
\mathcal{S}_b = \sum_{i=1}^{C} n_i (\vec{\mathcal{M}}_{(i)} - \vec{\mathcal{M}}) *_w (\vec{\mathcal{M}}_{(i)} - \vec{\mathcal{M}})^T,
$$

where $\mathcal{M}$ is the mean of all data samples and n_i is the number of samples in the class i.

The goal is to find a projection tensor $\mathcal{U}$ so as to maximize the between-class scatter while minimizing the within-class scatter (generally written as a scatter ratio). It can be show that the projection tensor in question can be computed by solving the generalized tensor eigenvalue problem as:

$$
(\mathcal{S}_w{}^{-1} *_w \mathcal{S}_b) *_w \mathcal{U} = \mathcal{D} *_w \mathcal{U}, \tag{7}
$$

where $\mathcal{U} = [\vec{\mathcal{U}}_{(1)}, \vec{\mathcal{U}}_{(2)}, \ldots, \vec{\mathcal{U}}_{(k)}] \in \mathbb{R}^{m \times k \times n}$ are the eigenmatrices corresponding to the k largest eigen-tuples of the diagonal tensor $\mathcal{D}$ using the tensor norm defined in **Definition 8** and the tensor inverse operation is outlined in **Definition 6**. Note that similar to its matrix counterpart, there are at most $c - 1$ nonzero eigentuples of (7), therefore the projection space has at most dimension $c - 1$. The projection tensor $\mathcal{U}$ can be obtained via the **t-eig$_w$** defined in Section 2.2. Finally, the tensor data $\mathcal{A} \in \mathbb{R}^{m \times \ell \times n}$ can be projected onto the new multilinear subspace $\mathcal{U} \in \mathbb{R}^{m \times k \times n}$ resulting in the new reduced feature tensor.

$$
\mathcal{B} = (\mathcal{U}^T *_w \mathcal{A}) \in \mathbb{R}^{k \times \ell \times n}. \tag{8}
$$

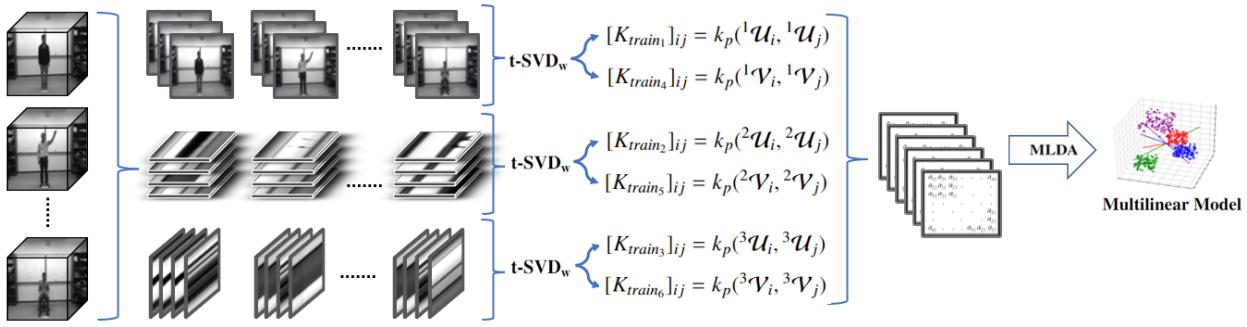


Figure 2: Graphical illustration of TGDA.

3.2.3. *Proposed Grassmann Tensor Discriminant Analysis*

In order to perform a discriminant analysis in the Grassmann manifold, we first map the tensor subspaces of the Grassmann manifold into a Hilbert space. Assume the third-order multilinear subspaces ${}^1\mathcal{U}_i, {}^2\mathcal{U}_i, {}^3\mathcal{U}_i, {}^1\mathcal{V}_i, {}^2\mathcal{V}_i, {}^3\mathcal{V}_i$ (for $i = 1 \cdots m$, m is the total number of samples in the training set) have already computed from the 3-mode **t-SVD$_w$** given in (4), (5), and (6). We can compute the gram matrices as:

$$
\begin{aligned}
[K_{train_1}]_{ij} &= k_p({}^1\mathcal{U}_i, {}^1\mathcal{U}_j),\\
[K_{train_2}]_{ij} &= k_p({}^2\mathcal{U}_i, {}^2\mathcal{U}_j),\\
[K_{train_3}]_{ij} &= k_p({}^3\mathcal{U}_i, {}^3\mathcal{U}_j),\\
[K_{train_4}]_{ij} &= k_p({}^1\mathcal{V}_i, {}^1\mathcal{V}_j),\\
[K_{train_5}]_{ij} &= k_p({}^2\mathcal{V}_i, {}^2\mathcal{V}_j),\\
[K_{train_6}]_{ij} &= k_p({}^3\mathcal{V}_i, {}^3\mathcal{V}_j),
\end{aligned}
$$

for all ${}^1\mathcal{U}_i, {}^2\mathcal{U}_i, {}^3\mathcal{U}_i, {}^1\mathcal{V}_i, {}^2\mathcal{V}_i, {}^3\mathcal{V}_i$ in the training set. We can construct our new data tensor $\mathcal{A}^\Phi \in \mathbb{R}^{m\times m\times 6}$ by stacking the gram matrices as frontal slices:

$$\mathcal{A}^\Phi(:,:,i) = K_{train_i},$$

for $i = 1, 2, \ldots, 6$.

The within-class scatter tensor can now be reformulated as:

$$\mathcal{S}_w^\Phi = \sum_{i=1}^{C} \mathcal{W}^{\Phi_i} *_w \mathcal{C}^{\Phi_i} *_w \mathcal{C}^{\Phi_i} *_w (\mathcal{W}^{\Phi_i})^T,$$

where $\mathcal{C}^{\Phi_i} \in \mathbb{R}^{k\times k\times 6}$ is the centering tensor of the i^{th} class which is an idempotent tensor defined in **Definition 9**.

$$\mathcal{C}^{\Phi_i} = \mathcal{C}^{\Phi_i} *_w \mathcal{C}^{\Phi_i}.$$

Therefore,

$$\mathcal{S}_w^\Phi = \sum_{i=1}^{C} \mathcal{W}^{\Phi_i} *_w \mathcal{C}^{\Phi_i} *_w (\mathcal{W}^{\Phi_i})^T,$$

where $\mathcal{W}^{\Phi_i} \in \mathbb{R}^{m\times k\times 6}$ is the kernel tensor, whose lateral slices are the lateral slices of the kernel tensor $\mathcal{A}^\Phi$ corresponding to i^{th} class. We also use a is a regularizer for making the computation stable:

$$\mathcal{S}_w^\Phi = \mathcal{S}_w^\Phi + \sigma\mathcal{I} \tag{9}$$

where $\mathcal{I} \in \mathbb{R}^{m\times m\times 6}$ is the identity tensor and σ is the regularization parameter.

Definition 11. The tensor $C^{\Phi} \in \mathbb{R}^{m \times m \times n}$ is called a centering tensor when right multiplied with a tensor $\mathcal{A} \in \mathbb{R}^{m \times m \times n}$ has the same effect as subtracting the mean of all lateral slices of the tensor and left multiplied with a tensor has the same effect as subtracting the mean of all horizontal slices of the tensor.

The mean of the lateral slices of the tensor $\mathcal{A}$ can be computed as:

$$\vec{\mathcal{M}} = \frac{1}{k} \sum_{i=1}^{k} \mathcal{A}^{(i)}.$$

To remove the mean $\vec{\mathcal{M}} \in \mathbb{R}^{k \times 1 \times n}$ from the tensor $\mathcal{A}$, we subtract the mean from the lateral slices of the tensor $\mathcal{A}$.

$$\bar{\mathcal{A}} = \mathcal{A} - \vec{\mathcal{M}}.$$

$C^{\Phi} \in \mathbb{R}^{k \times k \times n}$ is a centering tensor can be written as:

$$C^{\Phi} = \mathcal{I} - \frac{1}{k}\mathcal{J},$$
$$\mathcal{J} = \tilde{\mathcal{J}} \times_3 H_n^{-1},$$

where $\tilde{\mathcal{J}} \in \mathbb{R}^{k \times k \times n}$ is a tensor whose each entry is one, $\mathcal{I} \in \mathbb{R}^{k \times k \times n}$ is the identity tensor defined in **Definition 7** and H_n is the $n \times n$ Haar wavelet level-1 transformation matrix defined in Equation (1). We remind the reader that $\times_3$ is a mode-3 product and details can be found in [33, 17]. Multiplication by the centering tensor is a convenient analytical tool of removing the mean from a tensor.

$$\bar{\mathcal{A}} = \mathcal{A} *_w C^{\Phi}.$$

The fact that left multiplication removes the mean of the horizontal slices from the tensor can be similarly shown.

The between-class scatter tensor can be written as:

$$\mathcal{S}_b^{\Phi} = \sum_{i=1}^{C} (\vec{\mathcal{M}}^{\Phi_i} - \vec{\mathcal{M}}^{\Phi}) *_w (\vec{\mathcal{M}}^{\Phi_i} - \vec{\mathcal{M}}^{\Phi})^T, \tag{10}$$

where $\vec{\mathcal{M}}^{\Phi_i} \in \mathbb{R}^{m \times 1 \times 6}$ is the mean of the lateral slices of the i^{th} class of the kernel tensor $\mathcal{A}^{\Phi}$, whereas $\vec{\mathcal{M}}^{\Phi} \in \mathbb{R}^{m \times 1 \times 6}$ is the mean of the lateral slices of the tensor $\mathcal{A}^{\Phi}$.

The projection tensor $\mathcal{U}_w^{\Phi}$ can then be computed by solving the generalized tensor eigenvalue problem as:

$$(\mathcal{S}_w^{\Phi^{-1}} *_w \mathcal{S}_b^{\Phi}) *_w \mathcal{U}^{\Phi} = \mathcal{D}^{\Phi} *_w \mathcal{U}^{\Phi}, \tag{11}$$

where $\mathcal{U}^{\Phi} \in \mathbb{R}^{m \times c-1 \times 6}$ consists of eigenmatrices. Note that there are at most $c-1$ nonzero eigen-tuples of (11), therefore the projection space has at most dimension $c-1$.

Assume the multilinear subspaces of the testing inputs ${}^1\bar{\mathcal{U}}_i, {}^2\bar{\mathcal{U}}_i, {}^3\bar{\mathcal{U}}_i, {}^1\bar{\mathcal{V}}_i, {}^2\bar{\mathcal{V}}_i, {}^3\bar{\mathcal{V}}_i$ (for $i = 1 \cdots n$, n is the total number of samples in the testing set) have already computed from the 3-mode **t-SVD$_{\mathbf{w}}$**. Therefore, we can compute the gram matrices for testing set as:

$$[K_{test_1}]_{ij} = k_p({}^1\mathcal{U}_i, {}^1\bar{\mathcal{U}}_j),$$
$$[K_{test_2}]_{ij} = k_p({}^2\mathcal{U}_i, {}^2\bar{\mathcal{U}}_j),$$
$$[K_{test_3}]_{ij} = k_p({}^3\mathcal{U}_i, {}^3\bar{\mathcal{U}}_j),$$
$$[K_{test_4}]_{ij} = k_p({}^1\mathcal{V}_i, {}^1\bar{\mathcal{V}}_j),$$
$$[K_{test_5}]_{ij} = k_p({}^2\mathcal{V}_i, {}^2\bar{\mathcal{V}}_j),$$
$$[K_{test_6}]_{ij} = k_p({}^3\mathcal{V}_i, {}^3\bar{\mathcal{V}}_j),$$

We can construct our new data tensor $\mathcal{B}^{\Phi} \in \mathbb{R}^{m \times n \times 6}$ as:

$$B^{\Phi}(:,:,i) = K_{test_i},$$

for $i = 1, 2, \ldots, 6$. Finally, the training tensor data $\mathcal{A}^{\Phi} \in \mathbb{R}^{m \times m \times 6}$ and the testing training tensor data $\mathcal{B}^{\Phi} \in \mathbb{R}^{m \times n \times 6}$ can be projected onto the new multilinear subspace $\mathcal{U}^{\Phi} \in \mathbb{R}^{m \times c-1 \times 6}$ resulting in the new reduced feature tensor.

$$\bar{\mathcal{A}}^{\Phi} = \mathcal{U}^{\Phi^T} *_w \mathcal{A}^{\Phi} \in \mathbb{R}^{c-1 \times m \times 6}, \tag{12}$$

$$\bar{\mathcal{B}}^{\Phi} = \mathcal{U}^{\Phi^T} *_w \mathcal{B}^{\Phi} \in \mathbb{R}^{c-1 \times n \times 6}. \tag{13}$$

Once the projections $\bar{\mathcal{A}}^{\Phi}$ and $\bar{\mathcal{B}}^{\Phi}$ have been computed, classification is performed via nearest neighbor search for the closest match using the Frobenius norm. Graphical illustration of our proposed approach, namely tensor Grassmann discriminant analysis (TGDA)[2], is shown in Figure 2.

4. Experimental Results

In this section, we have presented experimental results on four well-known data sets: the Cambridge Hand Gesture [38], Weizmann [39], UTD-MHAD [40], and UCF sports action [41] data sets. We compare our proposed approach with the state-of-the-art algorithms to show the effectiveness of the proposed approach. The comprehensive explanation of each data set is given in the following.

4.1. Cambridge Hand-Gesture Database

Our first experiment is conducted using the Cambridge hand-gesture data set which has 900 video sequences of nine different hand gesture classes (100 video sequences per gesture class) [38]. The videos are collected from five different illumination sets labeled as Set1, Set2, Set3, Set4, and Set5. In the experiments, the same experimental setting in [42, 8, 10] is followed where the videos were converted to grayscale and resized to 20×20×20. Additionally, the videos from the Set5 was used for training and the videos from the Set1, Set2, Set3, and Set4 were used for testing.

In order to perform matrix-based methods GDA [4] and GGDA [5], each video sequence needs to be represented as a matrix. First, we row-scan the frames and create 400×20 matrices to represent the data set. The dimension of the rows is generally greater than the dimension of the columns due to the nature of row-scanning. As such, the matrix only spans 20 dimensions. For each sequence, a single subspace $U \in \mathbb{R}^{400 \times k}$ is produced by the singular value decomposition. Therefore, the maximum subspace dimension is $k = 20$. For our proposed approach, the subspaces ${}^1\mathcal{U}, {}^2\mathcal{U}, {}^3\mathcal{U}, {}^1\mathcal{V}, {}^2\mathcal{V}, {}^3\mathcal{V} \in \mathbb{R}^{20 \times 20 \times k}$ are third-order tensors and the maximum subspace dimension can be chosen as $k = 20$. Table 2 shows the experimental results of our methods compared with GDA [4] and GGDA [5] based on the subspace dimensions. The subspace dimension k is kept the same and 1-nearest neighbor classifier is performed for all methods to have a fair comparison. Additionally, the projection kernel is used for GDA and GGDA. The best classification accuracies of GDA and GGDA were obtained at subspace dimension 20, which was 82.1±5.9% and 86.7±3.9% respectively. Table 2 shows that our TDAG method is superior to matrix-based Grassmann discriminant analysis approaches. This experiment also indicates that using small subspace dimensions gives our proposed approach advantage over both GDA and GGDA. Table 3 illustrates the comparison of the proposed method with the other tensor-based methods, namely tensor canonical correlation anlysis (TCCA) [38], tangent bundles (TB) [7], product Grassmann manifold (PGM) [8], constrained multilinear discriminant analysis (CMDA) [43], direct general tensor discriminant analysis (DGTDA) [43], tensor-driven low-rank discriminant analysis (TLRDA) [44], and n-mode generalized difference subspace (n-mode GDS) [10] along with Grassmann kernels techniques, namely Grassmann discriminant analysis (GDA) [4], graph-embeded Grassmann discriminant analysis (GGDA) [5], and sparse-based classifier (SRC) [9].

[2]Our source code is available in the GitHub repository: https://github.com/Cagri-Ozdemir/TGDA

Table 2: Classification performance comparison between the proposed method and GDA [4] and GGDA [5] on the Cambridge Hand-Gesture data set based on subspace dimensions. The dimension of subspace is represented as k.

Test Set	Method	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
Set1	GDA	25.00%	76.11%	87.22%	91.67%	94.44%
	GGDA	27.22%	74.44%	85.00%	91.67%	92.78%
	TGDA	**98.33%**	**96.67%**	**97.22%**	**96.67%**	**95.56%**
Set2	GDA	11.11%	46.11%	59.44%	68.89%	71.11%
	GGDA	15.56%	40.56%	66.67%	77.78%	78.33%
	TGDA	**64.44%**	**86.67%**	**89.44%**	**92.22%**	**88.89%**
Set3	GDA	11.11%	41.67%	54.44%	67.78%	80.00%
	GGDA	16.11%	46.11%	71.67%	80.56%	81.11%
	TGDA	**77.22%**	**90.00%**	**93.33%**	**94.44%**	**92.22%**
Set4	GDA	11.67%	68.89%	75.00%	80.00%	86.67%
	GGDA	9.44%	73.89%	82.22%	81.11%	85.56%
	TGDA	**90.00%**	**91.67%**	**93.89%**	**93.33%**	**92.22%**
mean±std	GDA	14.11±5.45%	58.20±14.62%	60.03±12.96%	77.09±9.68%	83.06±8.58%
	GGDA	17.08±6.41%	58.75±15.54%	76.40±7.50%	82.78±5.29%	84.45±5.46%
	TGDA	**84.00±11.89%**	**91.25±3.61%**	**93.47±2.76%**	**94.17±1.65%**	**92.22±2.36%**

Table 3: Comparison of Classification Accuracy with State-of-the-Art Methods on the Cambridge Hand Gesture data set.

TCCA [42]	GDA [4]	TB [7]	GGDA [5]	PGM [8]	CMDA [43]	DGTDA [43]	TLRDA [45]	n-mode GDS [10]	SRC [9]	TGDA (Our method)
82±3.5%	82.1±5.9%	91±2.4%	86.7±3.9%	91.7±2.3%	42.78±25.84	36.67±19.29%	92.5±4.9%	93.5±2.1%	89.7±3.9%	**94.2±1.7%**

4.2. UTD-MHAD Database

The UTD-MHAD database contains 27 actions performed by 8 subjects [40]. The database contains RGB video, depth video, skeleton joint positions and inertial signals data. In our experiments, we only use RGB videos and follow the same experimental setting in [40] where the subject numbers 1, 3, 5, 7 were used for training, and the subject numbers 2, 4, 6, 8 were used for testing. The data set was pre-processed by using a people detector. We also grayscaled and resized each video to $20\times20\times20$. Table 4 illustrates the experimental results of our methods compared with GDA [4] and GGDA [5] based on the subspace dimensions. The subspace dimension k was kept the same and 1-nearest neighbor classifier was performed for all methods to have a fair comparison. Additionally, the projection kernel was used for GDA and GGDA. The best classification accuracies of GDA and GGDA were obtainted at the subspace dimension of 20 which are 76.62% and 72.92% respectively, whereas the best classification accuracy of the proposed approach was obtained at the subspace dimension of 6 which is 88.89%. In [40], the authors extract features from depth images and inertial sensors and combine the extracted features to achieve the best performance on their own data set, which is given in Table 5 as Kinect & Internal. In [46], motion and appearance features of RGB images are extracted using the histogram of Oriented Gradient (HOG) and Histogram of Optical Flow (HOF) descriptors. In [47], a new descriptor called 3D histograms of texture (3DHOT) has been introduced to extract discriminant features from depth images. Classification accuracies of these two descriptor-based techniques are also given in Table 5 as 3DHOT-MBC and STIP-BOW-SVM. Table 5 shows that the proposed approach appreciably better compared to given both tensor-based and descriptor-based methods.

Table 4: Classification performance comparison between the proposed method and GDA [4] and GGDA [5] on the UTD-MHAD data set based on subspace dimensions. The dimension of subspace is represented as k.

Method	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
GDA	20.83%	52.08%	59.72%	57.41%	62.96%	62.73%
GGDA	7.41%	44.21%	54.40%	61.34%	59.72%	62.27%
TGDA	**59.49%**	**78.94%**	**82.41%**	**83.80%**	**85.88%**	**88.89%**

Table 5: Comparison of Classification Accuracy with State-of-the-Art Methods on the UTD-MHAD data set.

GDA [4]	GGDA [5]	PGM [8]	CMDA [43]	DGTDA [43]	Kinect & Internal [40]	3DHOT-MBC [47]	STIP-BOW-SVM [46]	TGDA (Our method)
76.62%	72.92%	72.00%	56.05%	43.29%	79.1%	84.4%	67.37%	**88.89%**

4.3. Weizmann Database

The Weizmann database is a commonly used database for human action recognition [39]. The 90 videos coming from 10 categories of actions included bending (bend), jacking (jack), jumping (jump), jumping in places (pjump), running (run), gallopingside ways (side), skipping (skip), walking (walk), single-hand waving (wave1), and both-hands waving (wave2), which were performed by nine subjects. Our performance evaluation is based on the leave-one-person-out cross validation test. As such, as the actions were performed by nine subjects, we used each individual person as testing and the rest for training. The data set was pre-processed by using a people detector. We also grayscaled and resized each video to $20 \times 20 \times 20$. Table 6 shows the experimental results of our methods compared with GDA and GGDA based on the subspace dimensions. The best classification accuracies of GDA and TGDA were obtainted at the subspace dimension of 4 which are 93.33% and 94.44% respectively, whereas the best classification accuracy of GGDA approach was obtained at the subspace dimension of 5 which is 82.22%. Sparse tensor discriminant analysis (STDA) [48] and sparse tensor alignment (STA) [49] provide multilinear tensor extensions of manifold learning based algorithms to a sparse case. Table 7 illustrates that our proposed method performs better than STDA and STA techniques along with the other state-of-the-art methods.

Table 6: Classification performance comparison between the proposed method and GDA [4] and GGDA [5] on the Weizmann data set based on subspace dimensions. The dimension of subspace is represented as k.

Method	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
GDA	**78.89±12.86%**	75.56±14.99%	88.89±8.75%	93.33±8.17%	88.89±7.37%	**93.33±6.67%**
GGDA	41.11±7.37%	54.44±18.33%	75.55±14.23%	78.89±18.53%	82.22±10.30%	80.00±12.47%
TGDA	72.22±13.15%	**84.44±8.31%**	**91.11±3.14%**	**94.44±4.97%**	**91.11±5.67%**	91.11±5.67%

Table 7: Comparison of Classification Accuracy with State-of-the-Art Methods on the Weizmann data set.

GDA [4]	GGDA [5]	PGM [8]	CMDA [43]	DGTDA [43]	STDA [48]	STA [49]	TGDA (Our method)
93.33±8.17%	82.22±10.30%	75.56±10.66%	54.44±15.71	56.67%±10.54%	80.38±2.98%	79.33±3.54%	**94.44±4.97%**

4.4. UCF Sports Action Database

The UCF sports database [41] encompasses 10 sports actions recorded in real sport environment exhibiting the variations in background, illumination conditions, and occlusions, which make it a challenging data set. These actions include: golf swing, diving, lifting, kicking, running, riding horse, swing-bench, skateboarding, swing-side, and walking. The experimental results are based on Leave-One-Out (LOO) cross validation scheme. In LOO cross validation, all video sequences are used for training except one, which is used for testing the performance of the classifier. Frames in all video sequences were grayscaled and resized to 20×20. We use the region of interest provided with the data set. Table 8 shows the experimental results of our methods compared with GDA and GGDA based on the subspace dimensions. The best classification accuracies of GDA and GGDA were obtainted at the subspace dimension of 7 and 1 respectively, whereas the best classification accuracy of TGDA approach was obtained at the subspace dimension of 5. The HOG3D descriptor method [50] is based on histograms of 3D gradient orientations and provides promising classification rates for human action recognition problems; whereas RTW+eGDA [51] proposes a framework by extending the framework of GDA. Table 9 shows that our proposed method performs well compared to the aforementioned methods.

Table 8: Classification performance comparison between the proposed method and GDA [4] and GGDA [5] on the UCF sports data set based on subspace dimensions. The dimension of subspace is represented as k.

Method	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
GDA	60.66%	56.00%	62.00%	64.67%	68.00%	68.67%
GGDA	56.67%	40.00%	50.00%	46.00%	46.67%	48.67%
TGDA	**72.67%**	**80.00%**	**85.33%**	**85.33%**	**86.67%**	**86.67%**

Table 9: Comparison of Classification Accuracy with State-of-the-Art Methods on the UCF sports data set.

GDA [4]	**GGDA [5]**	**PGM [8]**	**CMDA [43]**	**DGTDA [43]**	**HOG3D [50]**	**RTW+eGDA [51]**	**TGDA (Our method)**
71.33%	56.67%	74.00%	48.67%	53.33%	85.6%	84.67%	**86.67%**

5. Conclusions and Feature Work

In this paper, we present a tensor-based discriminant analysis on the Grassmann manifold for human action recognition. We applied 3-mode **t-SVD$_{\mathbf{w}}$** and obtained third-order subspaces representing spatial and temporal information of a human action video. We also showed that an orthogonal tensor can be considered as a bundle of elements that naturally live on a Grassmann manifold. To apply the tensor-based discriminant analysis developed for Euclidean space, a novel tensor Grassmann projection kernel was also proposed to embed the Grassmann manifold into a Hilbert space, which satisfies necessary kernel validity conditions. Our experiments have demonstrated the superiority of our proposed approach over the state-of-the-art methods. Future work will be dedicated to evaluate a set of new positive definite kernels for third-order tensors.

6. Acknowledgments

The current research was supported in part by the Department of the Navy, Naval Engineering Education Consortium under Grant No. (N00174-19-1-0014) and the National Science Foundation under Grant No. (2007367).

References

[1] J. K. Aggarwal, M. S. Ryoo, Human activity analysis: A review, Acm Computing Surveys (Csur) 43 (3) (2011) 1–43.
[2] T. B. Moeslund, E. Granum, A survey of computer vision-based human motion capture, Computer vision and image understanding 81 (3) (2001) 231–268.
[3] C.-H. Kuo, R. Nevatia, How does person identity recognition help multi-person tracking?, in: CVPR 2011, IEEE, 2011, pp. 1217–1224.
[4] J. Hamm, D. D. Lee, Grassmann discriminant analysis: a unifying view on subspace-based learning, in: Proceedings of the 25th international conference on Machine learning, 2008, pp. 376–383.
[5] M. T. Harandi, C. Sanderson, S. Shirazi, B. C. Lovell, Graph embedding discriminant analysis on grassmannian manifolds for improved image set matching, in: CVPR 2011, IEEE, 2011, pp. 2705–2712.
[6] Y. M. Lui, J. R. Beveridge, M. Kirby, Action classification on product manifolds, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 833–839.
[7] Y. M. Lui, Tangent bundles on special manifolds for action recognition, IEEE Transactions on Circuits and Systems for Video Technology 22 (6) (2011) 930–942.
[8] Y. M. Lui, Human gesture recognition on product manifolds, The Journal of Machine Learning Research 13 (1) (2012) 3297–3321.
[9] K. Sharma, R. Rameshan, Image set classification using a distance-based kernel over affine grassmann manifold, IEEE Transactions on Neural Networks and Learning Systems 32 (3) (2020) 1082–1095.
[10] B. B. Gatto, E. M. dos Santos, A. L. Koerich, K. Fukui, W. S. Junior, Tensor analysis with n-mode generalized difference subspace, Expert Systems with Applications 171 (2021) 114559.
[11] M. T. Harandi, M. Salzmann, S. Jayasumana, R. Hartley, H. Li, Expanding the family of grassmannian kernels: An embedding perspective, in: European conference on computer vision, Springer, 2014, pp. 408–423.
[12] M. E. Kilmer, C. D. Martin, L. Perrone, A third-order generalization of the matrix SVD as a product of third-order tensors, Tufts University, Department of Computer Science, Tech. Rep. TR-2008-4.
[13] K. Braman, Third-order tensors as linear operators on a space of matrices, Linear Algebra and its Applications 433 (7) (Dec. 2010) 1241–1253.
[14] M. E. Kilmer, C. D. Martin, Factorization strategies for third-order tensors, Linear Algebra and its Applications 435 (3) (Aug. 2011) 641–658.
[15] H. Karner, J. Schneid, C. W. Ueberhuber, Spectral decomposition of real circulant matrices, Linear Algebra and Its Applications 367 (2003) 301–311.

[16] D. F. Gleich, C. Greif, J. M. Varah, The power and arnoldi methods in an algebra of circulants, Numerical Linear Algebra with Applications 20 (5) (Oct. 2013) 809–831.
[17] E. Kernfeld, M. Kilmer, S. Aeron, Tensor–tensor products with invertible linear transforms, Linear Algebra and its Applications 485 (2015) 545–570.
[18] D. A. Tarzanagh, G. Michailidis, Fast randomized algorithms for t-product based tensor operations and decompositions with applications to imaging data, SIAM Journal on Imaging Sciences 11 (4) (2018) 2629–2664.
[19] C. Ozdemir, R. C. Hoover and K. Caudle, Fast tensor singular value decomposition using the low-resolution features of tensors, in: 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, 2021, pp. 527–533.
[20] Z. Zhang, S. Aeron, Exact tensor completion using t-svd, IEEE Transactions on Signal Processing 65 (6) (2016) 1511–1526.
[21] P. Zhou, C. Lu, Z. Lin, C. Zhang, Tensor factorization for low-rank tensor completion, IEEE Transactions on Image Processing 27 (3) (2017) 1152–1163.
[22] S. Soltani, M. E. Kilmer, P. C. Hansen, A tensor-based dictionary learning approach to tomographic image reconstruction, BIT Numerical Mathematics 56 (4) (2016) 1425–1454.
[23] C. Zhang, W. Hu, T. Jin, Z. Mei, Nonlocal image denoising via adaptive tensor nuclear norm minimization, Neural Computing and Applications 29 (1) (2018) 3–19.
[24] M. E. Kilmer, K. Braman, N. Hao, R. C. Hoover, Third-order tensors as operators on matrices: A theoretical and computational framework with applications in imaging, SIAM Journal on Matrix Analysis and Applications 34 (1) (Feb. 2013) 148–172.
[25] N. Hao, M. E. Kilmer, K. Braman, R. C. Hoover, Facial recognition using tensor-tensor decompositions, SIAM Journal on Imaging Sciences 6 (1) (Feb. 2013) 437–463.
[26] R. C. Hoover, K. S. Braman, N. Hao, Pose estimation from a single image using tensor decomposition and an algebra of circulants, in: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, Sept. 2011, pp. 2928–2934.
[27] C. Ozdemir, R. C. Hoover, K. Caudle, 2DTPCA: A new framework for multilinear principal component analysis, in: 2021 IEEE International Conference on Image Processing (ICIP), IEEE, 2021, pp. 344–348.
[28] G. Strang, T. Nguyen, Wavelets and filter banks, SIAM, 1996.
[29] A. Jensen, A. la Cour-Harbo, Ripples in mathematics: the discrete wavelet transform, Springer Science & Business Media, June 2001.
[30] A. Haar, Zur theorie der orthogonalen funktionensysteme, Mathematische Annalen 69 (3) (1910) 331–371.
[31] I. Daubechies, Orthonormal bases of compactly supported wavelets ii. variations on a theme, SIAM Journal on Mathematical Analysis 24 (2) (Mar. 1993) 499–519.
[32] P. Porwik, A. Lisowska, The haar-wavelet transform in digital image processing: its status and achievements, Machine graphics and vision 13 (1/2) (Nov. 2004) 79–98.
[33] T. G. Kolda, B. W. Bader, Tensor decompositions and applications, SIAM review 51 (3) (2009) 455–500.
[34] R. C. Hoover, K. Caudle, K. Braman, Multilinear discriminant analysis through tensor-tensor eigendecomposition, in: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, Dec.2018, pp. 578–584.
[35] C. Ozdemir, R. C. Hoover, K. Caudle, K. Braman, High-order multilinear discriminant analysis via order-*n* tensor eigendecomposition, arXiv preprint arXiv:2205.09191.
[36] A. Edelman, T. A. Arias, S. T. Smith, The geometry of algorithms with orthogonality constraints, SIAM journal on Matrix Analysis and Applications 20 (2) (1998) 303–353.
[37] J. Hamm, D. Lee, Extended grassmann kernels for subspace-based learning, Advances in neural information processing systems 21.
[38] T.-K. Kim, S.-F. Wong, R. Cipolla, Tensor canonical correlation analysis for action classification, in: 2007 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2007, pp. 1–8.
[39] M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, in: Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, Vol. 2, IEEE, 2005, pp. 1395–1402.
[40] C. Chen, R. Jafari, N. Kehtarnavaz, Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor, in: 2015 IEEE International conference on image processing (ICIP), IEEE, 2015, pp. 168–172.
[41] M. Rodriguez, Spatio-temporal maximum average correlation height templates in action recognition and video summarization.
[42] T.-K. Kim, R. Cipolla, Canonical correlation analysis of video volume tensors for action categorization and detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (8) (2008) 1415–1428.
[43] Q. Li, D. Schonfeld, Multilinear discriminant analysis for higher-order tensor data classification, IEEE transactions on pattern analysis and machine intelligence 36 (12) (2014) 2524–2537.
[44] J. Zhang, C. Xu, P. Jing, C. Zhang, Y. Su, A tensor-driven temporal correlation model for video sequence classification, IEEE Signal Processing Letters 23 (9) (2016) 1246–1249.
[45] J. Zhang, Z. Li, P. Jing, Y. Liu, Y. Su, Tensor-driven low-rank discriminant analysis for image set classification, Multimedia Tools and Applications 78 (4) (2019) 4001–4020.
[46] A. B. Mahjoub, M. Atri, Human action recognition using rgb data, in: 2016 11th International Design & Test Symposium (IDT), IEEE, 2016, pp. 83–87.
[47] B. Zhang, Y. Yang, C. Chen, L. Yang, J. Han, L. Shao, Action recognition using 3d histograms of texture and a multi-class boosting classifier, IEEE Transactions on Image processing 26 (10) (2017) 4648–4660.
[48] Z. Lai, Y. Xu, J. Yang, J. Tang, D. Zhang, Sparse tensor discriminant analysis, IEEE transactions on Image processing 22 (10) (2013) 3904–3915.
[49] Z. Lai, W. K. Wong, Y. Xu, C. Zhao, M. Sun, Sparse alignment for robust tensor learning, IEEE transactions on neural networks and learning systems 25 (10) (2014) 1779–1792.
[50] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, C. Schmid, Evaluation of local spatio-temporal features for action recognition, in: Bmvc 2009-british machine vision conference, BMVA Press, 2009, pp. 124–1.
[51] L. S. Souza, B. B. Gatto, J.-H. Xue, K. Fukui, Enhanced grassmann discriminant analysis with randomized time warping for motion recognition, Pattern Recognition 97 (2020) 107028.