Spatio-Temporal Tensor Analysis for Whole-Brain fMRI Classification

Guixiang Ma[1*], Lifang He[2*], Chun-Ta Lu[1], Philip S. Yu[1,3], Linlin Shen[2†] and Ann B. Ragin[4]
[1]University of Illinois at Chicago, Chicago, IL, USA, {gma4, clu29}@uic.edu
[2]Shenzhen University, Shenzhen, China, {lifanghe, llshen}@szu.edu.cn
[3]Tsinghua University, Beijing, China, psyu@uic.edu
[4]Northwestern University, Chicago, IL, USA, ann-ragin@northwestern.edu

Abstract

Owing to prominence as a research and diagnostic tool in human brain mapping, whole-brain fMRI image analysis has been the focus of intense investigation. Conventionally, input fMRI brain images are converted into vectors or matrices and adapted in kernel based classifiers. fMRI data, however, are inherently coupled with sophisticated spatio-temporal tensor structure (*i.e.*, 3D space × time). Valuable structural information will be lost if the tensors are converted into vectors. Furthermore, time series fMRI data are noisy, involving time shift and low temporal resolution. To address these analytic challenges, more compact and discriminative representations for kernel modeling are needed. In this paper, we propose a novel spatio-temporal tensor kernel (STTK) approach for whole-brain fMRI image analysis. Specifically, we design a volumetric time series extraction approach to model the temporal data, and propose a spatio-temporal tensor based factorization for feature extraction. We further leverage the tensor structure to encode prior knowledge in the kernel. Extensive experiments using real-world datasets demonstrate that our proposed approach effectively boosts the fMRI classification performance in diverse brain disorders (*i.e.*, Alzheimer's disease, ADHD and HIV).

1 Introduction

Many neurological disorders (*e.g.*, Alzheimer's disease [1], neuro-AIDS [2]) are characterized in the early stages by latent ongoing brain injury. As a forefront Neuroimaging technique, functional Magnetic Resonance Imaging (fMRI) has been widely used for noninvasive interrogation of the brain. During the course of an fMRI experiment, a series of brain images are obtained in the resting state or while the subject performs a task tailored to activate a specific cognitive function. Over the last decade, machine learning classifiers, especially kernel method, *e.g.*, Support Vector Machines (SVM), have been successfully employed on fMRI images for analysis of neurological status and diagnosis [3, 4].

In this paper, we study the fMRI classification problem in the context of kernel modeling. Most work on fMRI classification focuses on analysis of specific brain regions of interest(ROI) [5]. However, ROI analysis is usually based on certain assumptions and may ignore additional valuable information in the image. Comparatively, whole-brain fMRI images provide comprehensive structural and functional information of the human brain, thus having higher exploratory power and lower bias [6, 7]. Typically, as shown in Fig. 1(a), a whole-brain fMRI image sample consists of a discrete time series of 3D image volumes (scans), where each volume consists of hundreds of thousands of voxels. Each voxel contains an intensity value that is proportional to the strength of the Nuclear Magnetic Resonance (NMR) signal emitted at the corresponding location in the brain volume [8]. Therefore, an fMRI brain image sample can be naturally represented as a fourth-order tensor with 3D space × time. How to appropriately utilize the information from such sophisticated spatio-temporal structure is a main issue in the kernel modeling task.

Most of conventional kernel methods convert a tensor to a vector (or a matrix), which is then adapted in the kernel modeling [9, 10]. However, voxels are often highly correlated with the surrounding voxels in the brain volume. For example, the adjacent voxels, marked with red and blue color in Fig. 1(b), exhibit similar patterns in Fig. 1(c). This kind of tensor-to-vector (or tensor-to-matrix) conversion would cause the loss of structural information such as the spatial arrangement of voxel-based features, particularly given that fMRI data has high spatial resolution [11].

A common solution is to focus on the 3D spatial domain of the fMRI brain images [7, 12]. For instance, in [12], the original 4D tensor of fMRI data is converted to a 3D tensor by averaging over the time dimension. Then the obtained 3D tensor is utilized in the kernel for

*Authors have equal contributions.
†Corresponding author.

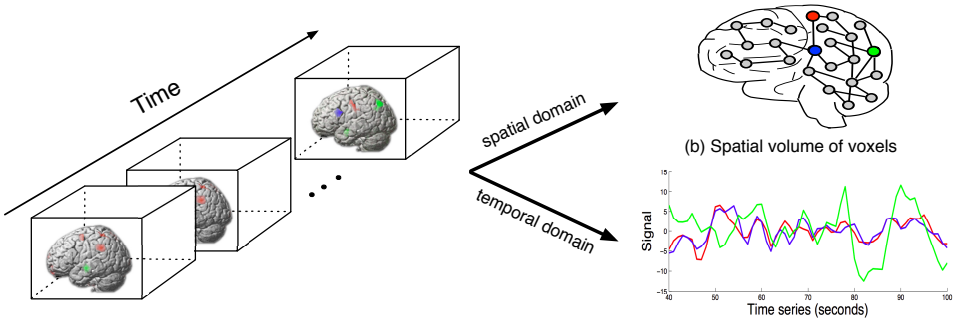


Figure 1: Example of fMRI brain images, which are inherently coupled with sophisticated spatio-temporal structure. Voxels are highly correlated with surrounding voxels in the spatial volume, and their signals are often very noisy in the time series.

classification. However, as shown in Fig. 1(c), the signal in each voxel of the brain volume changes along with time. If the time series is averaged, the varying trend in the time series that reflects the brain activity will be lost. Some studies [13, 14] have focused on analyzing the fMRI time series of each individual voxel while ignoring structural information in the spatial domain. For instance, the multilinear decomposition model [13] analyzes the time profile of the voxel vector converted from the 3D tensor in the spatial domain.

Although leveraging the spatio-temporal information is desired in building a predictive kernel method, it is very challenging due to the following three reasons:

Noisy fMRI time series analysis: Due to hardware reasons and subject factors (*e.g.*, thermal motion of electrons), there are often various nuisance components and random noise in fMRI signals, leading to a low signal-to-noise ratio (SNR) [11]. Since fMRI data has low temporal resolution, the signal of each voxel would not discriminatively change within a session of several time points, limiting ability to identify brain events in time frame. Furthermore, time shifts (delays), which occur naturally during the fMRI image acquisition process, should be taken into account while analyzing the data. How to filter the noise and extract discriminative information from the time series is critical in fMRI time series analysis.

Spatio-temporal feature extraction: Since fMRI data reflect brain activity from the spatial domain and temporal domain, a good feature extraction method should be able to extract a compact and informative representation from both domains while considering their correlations. Note that the time shift factor discussed previously should also be taken into consideration.

Kernel modeling: As discussed above, the existing works do not differentiate the spatial domain and temporal domain. How to incorporate both the correlation and the discrepancy between both domains into knowledge encoding is crucial for kernel modeling.

To deal with the above challenges, in this paper, we propose a Spatio-Temporal Tensor Kernel (STTK) framework for whole-brain fMRI image analysis. Specifically, we first perform time series extraction to reduce the noise and filter out the less informative time points in the original volumetric time series. Then we utilize the shifted CANDECOMP/PARAFAC (SCP) [15] factorization for feature extraction of the spatio-temporal data. Finally, spatio-temporal structure mapping is performed for kernel generation. Empirical studies on real-world resting-state fMRI brain images demonstrate that our proposed approach can significantly boost the fMRI classification performance on divergent disease diagnosis (*i.e.*, Alzheimer's disease, ADHD and HIV).

2 Preliminaries

In this section we define some necessary notions and notations related to tensors and then present the problem formulation. Before proceeding, we introduce some basic notations that will be used throughout this paper. Tensors (*i.e.*, multidimensional arrays) are denoted by calligraphic letters ($\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, $\cdots$), matrices by boldface capital letters ($\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\cdots$), vectors by boldface lowercase letters ($\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, $\cdots$), and scalars by lowercase letters (a, b, c, $\cdots$). The columns of a matrix are denoted by boldface lower letters with a subscript, *e.g.*, $\mathbf{a}_i$ is the ith column of matrix $\mathbf{A}$. The elements of a matrix or a tensor are denoted by lowercase letters with subscripts, *i.e.*, the $(i_1, \cdots, i_n)$ element of an n-th order tensor $\mathcal{A}$ is denoted by $a_{i_1,\ldots,i_n}$. $\mathbb{Z}^+$ is denoted by the set of positive integers. Additionally, we will often use Gothic letters ($\mathfrak{A}$, $\mathfrak{B}$, $\mathfrak{C}$, $\cdots$) to denote general sets or spaces, regardless of their specific nature.

2.1 Tensor Algebra

DEFINITION 1. (TENSOR) *An nth-order tensor is an element of the tensor product of n vector spaces, each of which has its own coordinate system.*

DEFINITION 2. (TENSOR PRODUCT) *Given order n and m tensors $\mathcal{A} \in \mathbb{R}^{I_1 \times \cdots \times I_n}$ and $\mathcal{B} \in \mathbb{R}^{I'_1 \times \cdots \times I'_m}$, their tensor product $\mathcal{A} \otimes \mathcal{B}$ is a tensor of order $n+m$ with the elements*

$$(\mathcal{A} \otimes \mathcal{B})_{i_1,\ldots,i_n,i'_1,\ldots,i'_m} = a_{i_1,\cdots,i_n} b_{i'_1,\cdots,i'_m} \tag{2.1}$$

Note that a rank-one tensor of order n is the tensor product of n vectors. Clearly, an important operation applicable to our analysis is the tensor product (also called the outer product). The tensor product generalizes from the Kronecker product, but results in another tensor rather than a block matrix, which naturally endows tensor with the structure of tensor product representations and tensor product spaces. The space is equipped with inner product and norm.

DEFINITION 3. (INNER PRODUCT) *The inner product of two same-sized tensors $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{I_1 \times \cdots \times I_n}$ is defined as the sum of the products of their elements:*

$$\langle \mathcal{A}, \mathcal{B} \rangle = \sum_{i_1=1}^{I_1} \cdots \sum_{i_n=1}^{I_n} a_{i_1,\ldots,i_n} b_{i_1,\ldots,i_n} \tag{2.2}$$

Clearly, for rank-one tensors $\mathcal{A} = \mathbf{a}^{(1)} \otimes \cdots \otimes \mathbf{a}^{(n)}$ and $\mathcal{B} = \mathbf{b}^{(1)} \otimes \cdots \otimes \mathbf{b}^{(n)}$, it holds that

$$\langle \mathcal{A}, \mathcal{B} \rangle = \langle \mathbf{a}^{(1)}, \mathbf{b}^{(1)} \rangle \cdots \langle \mathbf{a}^{(n)}, \mathbf{b}^{(n)} \rangle \tag{2.3}$$

For brevity, we denote $\mathbf{x}^{(1)} \otimes \cdots \otimes \mathbf{x}^{(m)}$ by $\prod_{i=1}^{m} \otimes \mathbf{x}^{(i)}$.

DEFINITION 4. (NORM) *The norm of a tensor $\mathcal{A}$ is defined to be the square root of the sum of all elements of the tensor squared, i.e.,*

$$\|\mathcal{A}\| = \sqrt{\langle \mathcal{A}, \mathcal{A} \rangle} = \sqrt{\sum_{i_1=1}^{I_1} \cdots \sum_{i_n=1}^{I_n} a_{i_1,\ldots,i_n} a_{i_1,\ldots,i_n}} \tag{2.4}$$

2.2 Problem Formulation In a typical fMRI classification task, we are given a collection of n training examples $\{\mathcal{X}_i, y_i\}_{i=1}^n \subset \mathfrak{X} \times \mathfrak{Y}$, where $\mathcal{X}_i \in \mathbb{R}^{I \times J \times K \times T}$ is the input fMRI sample with 3D space $\times$ time tensor form, and y_i is the class label of $\mathcal{X}_i$. The goal is to find a function $f : \mathfrak{X} \to \mathfrak{Y}$ that accurately predicts the label of an unseen example in $\mathfrak{X}$. In the kernel learning scenario, this problem can be formulated into the following optimization task:

$$f^* = \arg\min_{f \in \mathfrak{H}} \left(\frac{C}{n} \sum_{i=1}^{n} V\left(y_i, f(\mathcal{X}_i)\right) + \|f\|_{\mathfrak{H}}^2 \right), \tag{2.5}$$

where C controls the trade-off between the empirical risk and the regularization term $\|f\|_{\mathfrak{H}}^2$, $\mathfrak{H}$ is a set of functions forming a Hilbert space (the hypothesis space), and V is loss function that indicates how differences between y_i and $f(\mathcal{X}_i)$ should be penalized.

The attractiveness of kernel methods lies in its elegant treatment of nonlinear problems and its efficiency in high dimension. Different kernel methods or kernel machines arise from using different loss functions. In this paper, we use the hinge loss function $\max\{0, 1 - y_i f(\mathcal{X}_i)\}$ for support vector machine (SVM).

3 Kernel Modeling

Two components of kernel methods need to be distinguished: the kernel machine and the kernel function. The kernel machine encapsulates the learning task, which usually can be formulated as an optimization problem. The kernel function encapsulates the hypothesis language, *i.e.*, how to perform data transformation and knowledge encoding. By restricting to positive definite kernel functions, the optimization problem will be convex and solution will be unique. Throughout the paper, we take 'valid' to mean 'positive definite'.

DEFINITION 5. (POSITIVE DEFINITE KERNEL) *A symmetric function $\kappa : \mathfrak{X} \times \mathfrak{X} \to \mathbb{R}$ is a positive definite kernel on $\mathfrak{X}$ if, for all $n \in \mathbb{Z}^+$, $\mathcal{X}_1, \cdots, \mathcal{X}_n \in \mathfrak{X}$, and $c_1, \cdots, c_n \in \mathbb{R}$, it follows that $\sum_{i,j \in 1,\cdots,n} c_i c_j \kappa(\mathcal{X}_i, \mathcal{X}_j) \geq 0$.*

A kernel function κ corresponds to the inner product in some feature space (a Hilbert space), which is in general different from the representation space of the instances. The computational attractiveness of kernel methods comes from the fact that quite often a closed form of 'feature space inner products' exists [17]. Instead of mapping the data explicitly, the kernel can be calculated directly. According to *Mercer's theorem* [18], any valid kernel corresponds to an inner product in some feature space, and we can verify whether a kernel function is valid by the following Theorem [19].

THEOREM 1. *A function κ defined on $: \mathfrak{X} \times \mathfrak{X}$ is a positive definite kernel of $\mathfrak{H}$ if and only if there exists a feature mapping function $\phi : \mathfrak{X} \mapsto \mathfrak{H}$ such that*

$$\kappa(\mathcal{X}, \mathcal{Y}) = \langle \phi(\mathcal{X}), \phi(\mathcal{Y}) \rangle \tag{3.6}$$

for any $(\mathcal{X}, \mathcal{Y}) \in \mathfrak{X} \times \mathfrak{X}$.

In particular, an important property of positive definite kernels is that they are closed under sum, multiplication by a scalar and product [20].

By the *representer theorem* [21], the solutions of Eq. (2.5) can be given by

$$f^*(\mathcal{X}) = \sum_{i=1}^{n} c_i \kappa(\mathcal{X}_i, \mathcal{X}), \tag{3.7}$$

where $c_i \in \mathbb{R}$ are suitable coefficients, and κ is a valid kernel of $\mathfrak{H}$.

4 Spatio-Temporal Tensor Kernel framework

From the above discussions, it is clear that a good kernel should be data dependent. As noted in the introduction, fMRI data are inherently coupled with spatio-temporal tensor structure, involving time shift and have very low temporal resolution and SNR. To facilitate kernel learning for fMRI data, we propose a spatio-temporal tensor kernel (STTK) framework that takes both the correlation and discrepancy between spatial and temporal domains into account. This framework consists of three steps: (1) volumetric time series extraction for extracting discriminative information from the time series, (2) spatio-temporal feature extraction for obtaining a more compact and informative representation, and (3) tensor structure mapping for kernel generation.

4.1 Volumetric Time Series Extraction In fMRI time series extraction, a key issue is to determine the energy level for different time points. Most of existing work focus on single-voxel analysis [13], while they ignore the spatial correlations between voxels, which may lead to suboptimal outcomes. In this section we develop a volumetric time series extraction approach for fMRI time series. In particular, we show how the volumetric (spatial) correlations and the temporal varying properties can contribute to the energy levels.

Given an fMRI example $\mathcal{X} \in \mathbb{R}^{I\times J\times K\times T}$, let $\mathbf{x}_{i,j,k}[t] = \{\mathbf{x}_{i,j,k}, t = 1,\cdots,T\}$ be a T-element time series of voxel $x_{i,j,k}$, and $\mathcal{X}[t]$ is a volume of $\mathcal{X}$ at time point t. $\{E(t,\mathcal{X}[t]),\ E(t,x_{i,j,k,t})\}$ is the energy function of time point t, where E is separated by volume and voxel for computational purposes and $E(t,\mathcal{X}[t]) = \{E_{min}(t,\mathcal{X}[t]), E_{max}(t,\mathcal{X}[t])\}$ correspond to the *minima* and the *maxima* to be defined later.

The choice of energy function plays a critical role in explaining how the knowledge transforms into meanings and contexts. The success of time series extraction strongly depends on the data knowledge encoded into the energy function. Two important points must be emphasized. First, in order to reduce the noise present in the measurement, new features should be used to describe voxels, rather than using the noisy voxel intensities as features. Second, due to the low temporal resolution, each voxel signal would not experience a discriminative changing within a short measurement time period. It is necessary to make a discriminant analysis along time prior to the volume measurements. Based on these two points, we propose the following three-step procedure:

Voxel Energy Measurement: We first extract the *maxima* and *minima* (*extrema*) points for each voxel's time series using the extrema extraction method [22], which is an effective and efficient technique for single-voxel time series extraction and noise removal [14]. Let $\{(p_t, \mathbf{x}_{i,j,k}[p_t]), t = 1,\cdots,T_p\}$ and $\{(q_t, \mathbf{x}_{i,j,k}[q_t]), t = 1,\cdots,T_q\}$ be the *maxima* series and *minima* series of $\mathbf{x}_{i,j,k}[t]$, where p_t and q_t are the time indexes, T_p and T_q are the number of *maxima* and *minma*, repsectively. Then, for each voxel $x_{i,j,k,t}$, we measure its energy by

$$E(t,x_{i,j,k,t}) = \begin{cases} 1, & \text{if } t \in p_t \\ -1, & \text{if } t \in q_t \\ 0, & \text{otherwise} \end{cases} \tag{4.8}$$

where the values of 1 and -1 mean 'importance', and 0 means 'no importance'.

Volumetric Energy Measurement: We measure the energy of each volume by summing up the energies of all the voxels in it. In particular, we separately consider the maxima and minima voxels by

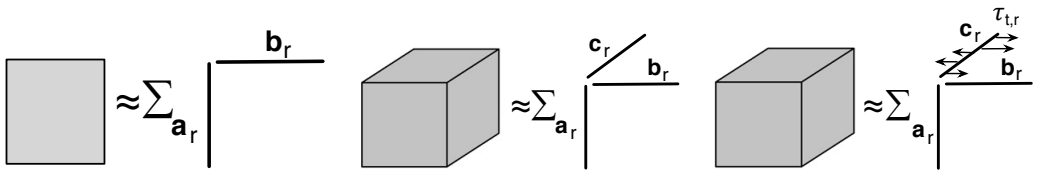
$$E_{max}(t,\mathcal{X}[t]) = \sum_{i,j,k} \max(E(t,x_{i,j,k,t}),0) \tag{4.9}$$

$$E_{min}(t,\mathcal{X}[t]) = \sum_{i,j,k} \max(-E(t,x_{i,j,k,t}),0) \tag{4.10}$$

Volumetric Time Series Extraction: We extract the time series from measured volumes based on E_{max} and E_{min}. Let ER_t be the time series extraction rate defined by N/T, where N is the number of extracted time points. Given an extraction rate ER_t, we first rank all the volumetric time points according to E_{max} and E_{min} respectively. Then we select the top-k time points from each of the two ranked time point sets and concatenate them, which forms the extracted time series, where k equals to $ER_t \times T/2$.

As an illustration, Fig. 2 shows the time series of a voxel with different time series extraction techniques. From the original time series (a), we can see that it is nontrivial to distinguish activation fluctuations from the background noise if no time series extraction is performed. Comparing with the time series (b) extracted using single-voxel technique, the time series (c), with the same amount of sampling time points as (b), extracted by our volumetric approach can better capture the significant changes of signal over time. For example, during the time interval $[70, 105]$ (between red lines), the signals in (c) experience notable irregular changes, which can also be observed from (a). Comparatively, (b) only captures the most distinct changes within this period. For the period $[0, 70]$, the original series shows slightly fluctuated changes, which can also be reflected by (c), while the time series (b) has much more changes.

This is majorly because the single-voxel technique chooses time points only based on the extrema of the single-voxel time series. In contrast, our approach



(a) Matrix Factorization (b) CP Factorization (c) Shifted CP Factorization

Figure 3: CP factorization is a generalization of matrix factorization to tensors. The SCP model allows shifts to occur over the second mode such that for each index of the third mode the component of the second mode is shifted a given amount.

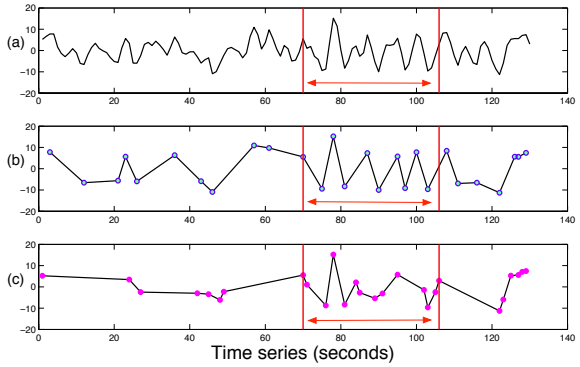


Figure 2: Illustration of the time series of a voxel with different time series extraction techniques. Each circle stands for an extracted time point. (a) is the original time series, (b) is the sequence extracted using single-voxel technique, where the time points with top 20% absolute values are extracted, and (c) is the sequence extracted using our approach, with $ER_t = 0.2$. Significant changes of signal occur in the interval between red lines.

performs the extraction based on the time varying volume series. Since the voxels of different regions in human brain are highly correlated and they usually collaboratively participate in a brain activity, their overall changing trend could better reflect the brain activity. By considering the time series of all the voxels in the volume, our extraction method incorporates both the spatial correlation of the volumetric voxels and the varying properties in the temporal domain into the analysis. Therefore, it can bring us more discriminative time series for fMRI brain image analysis.

4.2 Spatio-Temporal Feature Extraction Tensors provide a natural representation for fMRI data, but there is no guarantee that such representation will be good for kernel learning. From the characteristics of tensor, we know that the essential information in the tensor is embedded in its multi-way structure. Thus, one important aspect of kernel learning for such complex objects is to represent them by sets of key structural features which are easier to manipulate. Most of the previous work use CANDECOMP/PARAFAC (CP) factorization (as shown in Fig. 3) for fMRI data analysis, but it cannot well capture the structural information of a spatio-temporal tensor. Recently, it was found that shifted CP (SCP) factorization [13] is particularly effective for extracting such spatio-temporal structure. It can simultaneously consider the inter-mode correlations and the time shift in fMRI data, yielding a more compact representation of fMRI data. Motivated by these observations, we utilize SCP factorization to further perform feature extraction.

Given a tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K \times T}$, SCP factorizes it as

$$\mathcal{X} = \sum_{r=1}^{R} \mathbf{a}_r \otimes \mathbf{b}_r \otimes \mathbf{c}_r \otimes \mathbf{d}_r^{\tau} + \mathcal{E}, \tag{4.11}$$

where R is the rank of the tensor $\mathcal{X}$ defined as the smallest number of rank-one tensors in an exact SCP factorization, and the superscript τ denotes that the time shift will be along the forth mode (see Fig. 3 for an example), and $\mathcal{E}$ is the residual.

Remark that although SCP factorizes the data tensor, we can still recover the original data from the factorized results.

4.3 Tensor Structure Mapping Let us now consider how the above feature extraction results can be exploited to induce a kernel. Suppose we are given the SCP factorization of $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{I \times J \times K \times T}$ by $\mathcal{X} = \sum_{r=1}^{R} \mathbf{x}_r^{(1)} \otimes \mathbf{x}_r^{(2)} \otimes \mathbf{x}_r^{(3)} \otimes \mathbf{x}_r^{(4)\tau}$ and $\mathcal{Y} = \sum_{r=1}^{R} \mathbf{y}_r^{(1)} \otimes \mathbf{y}_r^{(2)} \otimes \mathbf{y}_r^{(3)} \otimes \mathbf{y}_r^{(4)\tau}$ respectively. We assume the tensor observations are mapped into the Hilbert space $\mathfrak{H}$ by

$$\phi : \mathcal{X} \to \phi(\mathcal{X}) \in \mathbb{R}^{H_1 \times H_2 \times H_3 \times H_4^{\tau}}. \tag{4.12}$$

Note that the projected tensor $\phi(\mathcal{X})$ has the same order with $\mathcal{X}$, but each mode dimension is higher and it is even an infinite dimension depending on the feature mapping function $\phi(.)$.

Based on the definition of the kernel function, it is easy to find that the feature space is a high-dimensional space of the original space, equipped with the same operations. Thus, we can factorize tensor data directly in the feature space in the same way as in the original space. This is formally equivalent to performing the following mapping:

(4.13)
$$\phi : \sum_{r=1}^{R} \prod_{i=1}^{3} \otimes \mathbf{x}_r^{(i)} \otimes \mathbf{x}_r^{(4)\tau} \rightarrow \sum_{r=1}^{R} \prod_{i=1}^{3} \otimes \phi(\mathbf{x}_r^{(i)}) \otimes \phi(\mathbf{x}_r^{(4)\tau}).$$

In this sense, it corresponds to mapping tensors into high-dimensional tensors that retain the original structure. More precisely, it can be regarded as mapping the original data into tensor feature space and then conducting the SCP factorization in the feature space.

After mapping the SCP factorization of the data into the tensor product feature space, the kernel can be defined directly with the inner product in that feature space. Thus, we derive our STTK:

$$\begin{aligned} &\kappa\left(\sum_{r=1}^{R} \prod_{i=1}^{3} \otimes \mathbf{x}_r^{(i)} \otimes \mathbf{x}_r^{(4)\tau}, \sum_{r=1}^{R} \prod_{i=1}^{3} \otimes \mathbf{y}_r^{(i)} \otimes \mathbf{y}_r^{(4)\tau}\right) \\ &= \sum_{p=1}^{R} \sum_{q=1}^{R} \prod_{i=1}^{3} \kappa\left(\mathbf{x}_p^{(i)}, \mathbf{y}_q^{(i)}\right) \kappa\left(\mathbf{x}_p^{(4)\tau}, \mathbf{y}_q^{(4)\tau}\right). \end{aligned} \tag{4.14}$$

From its derivation, we know such a kernel can take the multi-way spatio-temporal structure flexibility into account. In general, the STTK is an extension of the conventional kernels in the vector space to tensor space, and each vector kernel can be used in this framework for fMRI classification analysis in conjunction with kernel machines. Our positive result can be viewed as saying that designing a good tensor kernel function is much like designing a good tensor structure in the feature space.

5 Experiments and Evaluation

In order to empirically evaluate the effectiveness of the proposed approach for fMRI classification, we test our model on real fMRI data and compare with several state-of-the-art kernel methods in fMRI study.

5.1 Data Collection and Preprocessing In this work, we consider three real resting-state whole-brain fMRI image datasets as follows:

- *Alzheimer's Disease (ADNI)*[1]: It contains fMRI images of 33 subjects, each with a series of $61 \times 73 \times 61$ scans for 130 time points. These subjects are AD patients (positive) or normal people (negative).
- *Human Immunodeficiency Virus Infection (HIV)* [23]: This dataset contains fMRI brain images of 83 subjects, each with a series of $61 \times 73 \times 61$ scans for 255 time points. These subjects are early HIV patients (positive) or normal controls (negative).
- *Attention Deficit Hyperactivity Disorder (ADHD)*[2]: This dataset contains the resting-state fMRI images of 100 subjects, each with a series of scans for $58 \times 49 \times 47$ voxels. Subjects are either ADHD patients (positive) or normal controls (negative). Different from previous datasets, the lengths of time series for different subjects in ADHD dataset are not the same, ranging from 74 to 257.

In the derived datasets, each 3D fMRI scan has the NIFTI format. We convert each scan to a 3D tensor using SPM8[3]. Then we use SPM8 toolbox to preprocess these data, including images realignment, slice timing correction and normalization. We also perform spatial smoothing on these functional images with an 8mm FWHM Gaussian kernel for increasing signal-to-noise ratio (SNR). REST[4] is used afterwards for band-pass filtering (0.01-0.08 Hz) and linear trend removing of the time series.

5.2 Baselines and Metrics In order to establish a comparative study, we use five kernel learning methods as baselines. We use the classification accuracy as the evaluation metric.

- **Factor kernel (FK)** [9]: a matrix unfolding based tensor kernel. The constituent kernels are from the class of Gaussian RBF kernels.
- **sKL** [10]: a kernel defined based on the symmetric Kullback-Leibler divergence, where the tensors are also unfolded into matrices, which has been applied to reconstruct 3D movement.
- **DuSK** [12]: a tensor kernel based upon CP factorization. The authors average the fMRI data over the temporal dimension and apply DuSK on the obtained 3D fMRI data. For evaluation, we implement DuSK in both the 3D spatial data setting and the 4D spatio-temporal data setting, which are denoted as **S-DuSK** and **ST-DuSK**, respectively.
- **STTK:** our proposed spatio-temporal tensor kernel. To evaluate the effectiveness of volumetric time series extraction, we employ STTK with and without volumetric time series extraction and denote them as **STTK** and $\mathbf{STTK}_{nonTE}$ respectively. Specifically, to study the importance of temporal correlations of the fMRI brain images within the time series, we randomly permute the order of the time dimension of the fMRI data, and then apply our STTK to it. We denote this case as $\mathbf{STTK}_{permT}$.

We apply each kernel learning method in SVM and evaluate their performance. Specifically, we apply all

[1] http://adni.loni.usc.edu/
[2] http://neurobureau.projects.nitrc.org/ADHD200/
[3] http://www.l.ion.uc.ac.uk/spm/software/spm8
[4] http://resting-fmri.sourceforge.net

Table 1: Summary of compared methods. ST means Spatio-Temporal, C is the trade-off parameter, σ is the kernel width parameter, R is the rank of tensor factorization, and ER_t is the time series extraction rate.

Property	FK [9]	sKL [10]	S-DuSK [12]	ST-DuSK [12]	STTK$_{nonTE}$	STTK
Type of Input Data	Unfolded Matrices	Unfolded Matrices	Spatial Tensor	ST Tensor	ST Tensor	ST Tensor
Type of ST Correlation Exploited	One-way	One-way	Three-way	Multi-way	Multi-way	Multi-way
Differentiating Space V.S. Time	No	No	No	No	Yes	Yes
Time Series Feature Extraction	No	No	No	No	No	Yes
Parameters	C, σ	C, σ, R	C, σ, R	C, σ, R	C, σ, R	C, σ, R, ER_t

the six methods on ADNI and HIV datasets. For the ADHD dataset, the lengths of the time series are different for different subjects, while Factor kernel, sKL, ST-DuSK and STTK$_{nonTE}$ require dimensions of different samples must agree. Thus, we only apply S-DuSK, STTK$_{permT}$ and STTK on ADHD dataset. We use LibSVM [24], a widely used implementation of SVM, with Gaussian RBF kernel as the classifier. Table 1 summarized the compared methods. The optimal trade-off parameter for all the methods is selected from $C \in \{2^{-5}, 2^{-4}, \ldots, 2^5\}$, the kernel width parameter is selected from $\sigma \in \{2^{-5}, 2^{-4}, \ldots, 2^5\}$, the optimal rank R is determined by grid search from $\{1, 2, \ldots, 8\}$, and the time series extraction rate ER_t is chosen from $[0, 1]$. Here we set the time series extraction rate ER_t to be 0.2, *i.e.*, only 20% of the time sequences will be kept. In the experiment, 5-fold cross validations are performed. We repeated this process for 50 times and report the average classification accuracy as the result.

5.3 Classification Performance As shown in Table 2, our STTK method performs the best on all three datasets in terms of classification accuracy. Among the listed kernel methods, Factor kernel and sKL unfold the original tensor data into matrices while all the other methods preserve the spatial tensor structure during the learning process. As can be seen from the results, Factor kernel and sKL achieve a relatively lower accuracy on both the ADNI dataset and HIV dataset. This implies that unfolding tensor into matrices would lose the spatial structural information, leading to the degraded performance. Another observation is that DuSK achieves a quite high accuracy when applied in the three-dimensional spatial data setting, while the accuracy decreases to a great extent on the four-dimensional spatio-temporal fMRI data. This is majorly due to the fact that the time series of fMRI data are very noisy, involving time shift and with low SNR. Extending DuSK to the spatio-temporal domain without proper treatments would even damage its performance.

Comparatively, our proposed STTK properly encodes the prior knowledge of time series analysis with the spatio- temporal structural information into one tensor based kernel model. Therefore, the classification accuracy of STTK is much higher than that of ST-DuSK, especially on the ADNI dataset. Furthermore,

Table 2: Classification accuracy comparison (mean ± standard deviation)

	ADNI	HIV	ADHD
FK	0.593 ± 0.029	0.663 ± 0.011	N/A
sKL	0.510 ± 0.030	0.645 ± 0.021	N/A
S-DuSK	0.731 ± 0.021	0.718 ± 0.005	0.622 ± 0.010
ST-DuSK	0.576 ± 0.052	0.642 ± 0.023	N/A
STTK$_{nonTE}$	0.710 ± 0.010	0.693 ± 0.006	N/A
STTK$_{permT}$	0.583 ± 0.020	0.615 ± 0.021	0.594 ± 0.018
STTK	$\mathbf{0.759 \pm 0.022}$	$\mathbf{0.762 \pm 0.010}$	$\mathbf{0.680 \pm 0.013}$

by extracting the most significant features in the time series at an appropriate compression rate, our STTK can better discriminate the fMRI patterns with different medical status. Meanwhile, this volumetric time series extraction strategy enables us to analyze fMRI time series with different lengths (*e.g.*, the ADHD dataset used in the experiment). The experimental results demonstrate the effectiveness and considerable advantages of our proposed methods in the fMRI study.

As can be seen in Table 2, another notable result is that STTK$_{permT}$ achieves a much lower accuracy than STTK, which means the random permutation of the temporal sequential order of fMRI brain images degrades the classification performance. This implies that the temporal order of the fMRI brain images is very important for the classification. This is mainly because that the original varying trend of the fMRI time series reflects the sequential brain activity within the period. If the temporal order of the fMRI brain images is permuted, the original temporal correlation of the fMRI brain images would be damaged. This result also demonstrates that our STTK method captures the temporal correlation of fMRI data well.

5.4 Parameter Sensitivity Although the optimal values of the parameters in our proposed STTK are found using cross-validation, it is of interest to see the sensitivity of STTK to the time series extraction rate ER_t and the rank of tensor factorization R.

We first evaluate the classification performance of STTK with varying ER_t. We vary ER_t from 0.1 to 1.0 on ADNI and HIV datasets. For ADHD dataset, the lengths of time series for different subjects are quite different, varying from 74 to 257. We extract the same number of time points from each of them, and then compute the average extraction rate, and use it as the extraction rate for ADHD dataset. Since the

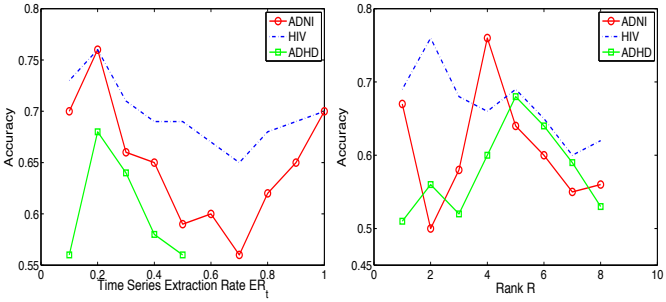


Figure 4: Parameter sensitivity

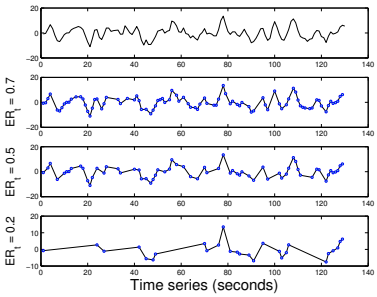


Figure 5: Time series of a voxel extracted with varying time series extraction rate ER_t.

average extraction rate reaches its maximum around 0.58 due to the different lengths of the time series, here we vary ER_t from 0.1 to 0.5 for the evaluation on ADHD dataset. As shown in Fig. 4, the value of ER_t significantly impacts the classification accuracy. We can find that the accuracy declines when $ER_t > 0.2$. This indicates, counterintuitively, keeping more time points (with higher ER_t) does not improve the accuracy; instead, it may even lead to a worse performance. As illustrated in Fig. 5, the time series extracted with $ER_t = 0.5$ and the one extracted with $ER_t = 0.7$ contain many redundant time points, especially in the time interval [0, 70], which may degrade the performance. Although keeping even more time points might be helpful, as the accuracy starts to increase when $ER_t > 0.7$, we can notice that the optimal results for all datasets are achieved when $ER_t = 0.2$. This reflects the fact that fMRI time series are commonly noisy, containing many redundant time points that are insignificant for disease diagnosis. With an appropriate value of ER_t, the volumetric time series extraction enables STTK to greatly filter the background noise, while preserving the most discriminative patterns in the fMRI time series.

Next, we evaluate the sensitivity of STTK to the rank R of tensor factorization. We fix ER_t at 0.2 which is the optimal value for each dataset, and vary R from 1 to 8 with a step size of 1. As shown in Fig. 4, the rank parameter R has a significant effect on the classification accuracy and the optimal value of R depends on the datasets. In general, the optimal value of R lies in the range between 2 and 5, which may provide a good guidance for selection of the R in advance. How to determine the optimal rank for a specific tensor factorization method is beyond the scope of this paper and still remains an open research problem [25, 26].

6 Related Work

Our work relates to a vast literature on spatio-temporal data analysis, tensor analysis techniques, and kernel learning. We present a selection of such works below.

Spatio-Temporal Data Analysis: Spatio-temporal data analysis has attracted considerable attention recently. Many models have been conducted to address the challenges in different contexts [27]. However, these models usually require domain knowledge since they make strong assumptions on the spatial and temporal correlation of the data. Some models have been used in the spatio-temporal fMRI brain image analysis [28], while most of them treat spatial domain and temporal domain separately. For instance, in [29], spatial analysis is performed via general linear modelling (GLM), while temporal analysis is done with a direct comparison of BOLD response estimates between regions.

Tensor Factorizations: Our work is also motivated by recent advances in tensor factorization and its applications in the fMRI data analysis [30]. A comprehensive survey on tensor factorization can be found in [16]. One of the most commonly used one is CP factorization. In the spatio-temporal tensor setting, the shifted CP is more frequently used [13], but for exploratory analysis. In this study, we employ it to facilitate kernel learning.

Kernel learning: Several tensor based kernel methods have been recently investigated [9, 10, 17]. Most of them focus on learning kernel via matrix unfolding, thus only capturing the one-way relationship within the tensor data. The multi-way structures within tensor data are already lost before the kernel construction. The problem of how to build kernel directly on tensor data has not been well studied. A first attempt in this direction is related to CP factorization proposed in [12], while it has the same drawback as CP factorization.

7 Conclusion

In this paper, we have introduced a spatio-temporal tensor kernel (STTK) modeling method, with an applica-

tion to whole-brain fMRI classification. Different from conventional kernel methods, our approach exploits the inherent spatio-temporal structure to facilitate kernel learning, while considering both the correlation and discrepancy between the spatial domain and the temporal domain. STTK consists of three steps: (1) volumetric time series extraction for extracting discriminate information from the time series, (2) spatio-temporal feature extraction for obtaining a more compact and informative representation, and (3) tensor structure mapping for kernel generation. Empirical studies on real-world fMRI brain images demonstrate that our approach can significantly boost the fMRI classification performance in three different brain disorders (*i.e.*, Alzheimer's disease, ADHD and HIV).

Acknowledgments

This work is supported in part by NSF (III-1526499, CNS-1115234, OISE-1129076), NSFC (61272050, 61273295, 61472089, 61503253), NSFC-Guangdong Joint Found (U1501254), Google Research Award, the Science Foundation of Guangdong Province (2014A030313556), and the National Institutes of Health (R01-MH080636).

References

[1] J. Ye, K. Chen, T. Wu, et al. Heterogeneous data fusion for alzheimer's disease study. *ACM SIGKDD*, 2008.

[2] A. Ragin, H. Du, R. Ochs, et al. Structural brain alterations can be detected early in HIV infection. *Neurology*, 79(24): 2328-2334, 2012.

[3] W. Koch, S. Teipel, S. Mueller, et al. Diagnostic power of default mode network resting state fMRI in the detection of Alzheimer's disease. *Neurobiology of Aging*, 33(3): 466-478, 2012.

[4] P. Matthews, G. Honey, and E. Bullmore. Applications of fMRI in translational medicine and clinical practice. *Nature Reviews Neuroscience*, 7(9): 732-744, 2006.

[5] M. McKeown, J. Li, X. Huang, et al. Local linear discriminant analysis (LLDA) for group and region of interest (ROI)-based fMRI analysis. *Neuroimage*, 37(3): 855-865, 2007.

[6] C. Ecker, V. Rocha-Rego, P. Johnston, et al. Investigating the predictive value of whole-brain structural MR scans in autism: a pattern classification approach. *Neuroimage*, 49(1): 44-56, 2010.

[7] X. Song, L. Meng, Q. Shi, et al. Learning Tensor-Based Features for Whole-Brain fMRI Classification. *MICCAI*, 2015.

[8] R. Graaf and K. Kevin. Methods and apparatus for compensating field inhomogeneities in magnetic resonance studies. US Patent No. 8035387, 2011.

[9] M. Signoretto, L.D. Lathauwer, and JAK Suykens. A kernel-based framework to tensorial data analysis. *Neural Networks*, 24(8): 861-874, 2011.

[10] Q. Zhao, G. Zhou, T. Adali, et al. Kernelization of tensor-based models for multiway data analysis: Processing of multidimensional structured data. IEEE Signal Processing Magazine, 30(4): 137-148, 2013.

[11] M. Lindquist. The statistical analysis of fMRI data. *Statistical Science*, 23(4): 439-464, 2008.

[12] L. He, X. Kong, P.S. Yu, et al. Dusk: A dual structure-preserving kernel for supervised tensor learning with applications to neuroimages. *SDM*, 2014.

[13] M. Mørup, L. Hansen, S. Arnfred, et al. Shift-invariant multilinear decomposition of neuroimaging data. *Neuroimage*, 42(4): 1439-1450, 2008.

[14] F. Deng, D. Zhu, J. Lv, et al. FMRI signal analysis using empirical mean curve decomposition. *IEEE TBME*, 60(1): 42-54, 2013.

[15] R. Harshman, S. Hong, and M .Lundy. Shifted factor analysis–Part I: Models and properties. *Journal of Chemometrics*, 17(7): 363-378, 2003.

[16] T. Kolda and B. Bader. Tensor decompositions and applications. *SIAM review*, 51(3): 455-500, 2009.

[17] T. Gärtner. A survey of kernels for structured data. *ACM SIGKDD Explorations Newsletter*, 5(1): 49-58, 2003.

[18] V. Vapnik. The nature of statistical learning theory. *Springer Science & Business Media*, 2013.

[19] A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*, 2004.

[20] N. Cristianini and J. Shawe-Taylor. An introduction to support vector machines and other kernel-based learning methods. *Cambridge university press*, 2000.

[21] B. Schölkopf, R. Herbrich, and A. Smola. A generalized representer theorem. *Computational Learning Theory*, pp. 416-426, 2001.

[22] E. Fink and H. Gandhi. Compression of time series by extracting major extrema. *JETAI*, 2011.

[23] X. Wang, P. Foryt, R. Ochs, et al. Abnormalities in resting-state functional connectivity in early human immunodeficiency virus infection. *Brain Connectivity*, 1(3): 207-217, 2011.

[24] C. Chang and C. Lin. LIBSVM: a library for support vector machines. *ACM TIST*, 2(3): 27, 2011.

[25] D. Vin and L-H Lim. Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIMAX*, 30(3): 1084-1127, 2008.

[26] Z. Hao, L. He, B. Chen, et al. A linear support higher-order tensor machine for classification. *IEEE TIP*, 22(7): 2911-2920, 2013.

[27] N. Cressie and K. Christopher. Statistics for spatio-temporal data. *Wiley*, 2011.

[28] V. Oikonomou, K. Blekas, and L. Astrakas. A sparse and spatially constrained generative regression model for fMRI data analysis. *IEEE TBME*, 2012.

[29] S. Haller, M. Klarhoefer, J. Schwarzbach, et al. Spatial and temporal analysis of fMRI data on word and sentence reading. *European Journal of Neuroscience*, 26(7): 2074-2084, 2007.

[30] L. Kuang, Q. Lin, X. Gong, et al. Multi-subject fMRI data analysis: Shift-invariant tensor factorization vs. group independent component analysis. *ChinaSIP*, 2013.