Mach Learn
DOI 10.1007/s10994-009-5159-x

Decomposing the tensor kernel support vector machine for neuroscience data with structured labels

David R. Hardoon · John Shawe-Taylor

Received: 27 February 2009 / Revised: 16 September 2009 / Accepted: 22 October 2009

Abstract The tensor kernel has been used across the machine learning literature for a number of purposes and applications, due to its ability to incorporate samples from multiple sources into a joint kernel defined feature space. Despite these uses, there have been no attempts made towards investigating the resulting tensor weight in respect to the contribution of the individual tensor sources. Motivated by the increase in the current availability of Neuroscience data, specifically for two-source analyses, we propose a novel approach for decomposing the resulting tensor weight into its two components without accessing the feature space. We demonstrate our method and give experimental results on paired fMRI image-stimuli data.

Keywords Tensor kernel · Support vector machine · Decomposition · fMRI

1 Introduction

Recently, machine learning methodologies have been increasingly used to analyse the relationship between stimulus categories and neuroscience data, such as functional Magnetic Resonance Imaging (fMRI) responses (Carlson et al. 2003; Mitchell et al. 2004; LaConte et al. 2005; Mourão-Miranda et al. 2005). Furthermore, due to recent improvements in neuroscience scanning technology, there has been an increased interest in analysing various conditions using cross-domain multiple sources (e.g., medical devices); such as learning the relationship between brain structure and genetic representation (Hardoon et al. 2009), as well as conducting experiments aimed at analysing brain responses to (complex) structured data such as; listening to music (Koelsch et al. 2006), watching a movie (Anderson

Editors: Nicolo Cesa-Bianchi and Gayle Leen.

D.R. Hardoon (✉) · J. Shawe-Taylor
Centre for Computational Statistics and Machine Learning, Department of Computer Science, University College London, London, UK
e-mail: D.Hardoon@cs.ucl.ac.uk

J. Shawe-Taylor
e-mail: jst@cs.ucl.ac.uk

Published online: 09 December 2009

et al. 2006), or observing real-world images (Mourão-Miranda et al. 2006). In this paper we focus on supervised learning with multiple sources and propose a novel methodology for decomposing source-independent weight vectors from the joint kernel defined feature space.

Current analysis techniques conventionally rely on replacing the structured stimulus with a simple categorical representation for the stimulus type (e.g. type of task). This representation does not enable the analysis to take into account the information that can potentially be gained from the structure of the stimuli. In a recent study Hardoon et al. (2007) proposed extending the standard methodology by incorporating the structural information of image-stimuli into the brain analysis via Canonical Correlation Analysis (CCA) as an unsupervised fMRI analysis technique. In this study, the simple categorical description of the stimulus type (e.g., $+1/-1$) was replaced by a more informative vector of stimulus features. The unsupervised procedure in Hardoon et al. (2007) was indeed able to discriminate between the tasks as well as providing an insight into the corresponding feature representations for both the brain and image data, i.e. producing a heat map of brain region activity as well as an indication of the image features corresponding to respective tasks. Interpreting the weight vectors as brain activity maps has been proposed by Mourão-Miranda et al. (2005) who applied the SVM to map whole fMRI volumes (brain states) from different subjects to different classes without prior selection of features. They demonstrated that the SVM produces spatial maps, i.e. the weight vector that were robust and comparable to the GLM (Friston et al. 1995) standard fMRI analysis.

A potential disadvantage of the CCA technique (Hardoon et al. 2007) is that it performs feature selection in an unsupervised way (O'Toole et al. 2007). The discriminative power, and the resulting brain analysis, may not conform to the original goal of the experiment and therefore would potentially not be of interest to the neuroscientist. For example, analysing the cognitive response for structured images containing natural scenery (with overlapping features) may focus attention on processing of natural scenery rather than features relevant to the experimental design/interest. This potential disadvantage of an unsupervised analysis may pose a "high risk" given the cost and time required to conduct such experiments. Therefore, we propose to explicitly use the label/task information while learning the relationship between the brain and structured stimulus by using a Support Vector Machine (SVM) with a tensor kernel. The tensor product kernel allows the combination of several sources into the kernel defined feature space (we briefly review the tensor product kernel in Sect. 2). We subsequently propose a novel decomposition of the resulting tensor weight vector into individual weight components without accessing the two feature spaces.

Tensor product kernels have been used across the machine learning literature for various purposes. We briefly review a number of these uses, for example; Szedmak et al. (2005) has used the tensor product kernel to construct a new maximum margin framework for multi-class and multi-view learning at a one-class complexity and Szedmak et al. (2007) has recently shown how the maximum margin framework relates to CCA, a powerful unsupervised tool for discovering relationships between different sources of information. Kondor and Lafferty (2002) proposed a general method of constructing natural families of kernels over discrete structures, based on the matrix exponentiation idea, while Ben-Hur and Noble (2005), Martin et al. (2005), Qiu and Noble (2008) explored predicting edges in a protein interaction or co-complex network, using a tensor product transformation to derive a kernel on protein pairs from a kernel on individual proteins. Furthermore, Weston et al. (2007) have used tensor kernels to explore methodologies for solving high dimensional estimation problems between pairs of arbitrary data types as a regression problem.

Despite the many uses of tensor kernels across the machine learning literature we have not found any attempts made towards representing the resulting weight vector, i.e. when

two sources[1] have been used to construct the hyper-plane the resulting weight vector does not elucidate the respective contributions of each of the sources. We are motivated by the specific problem when fMRI and complex image stimuli are used within a tensor kernel SVM. In this problem, further to achieving good results, we wish to obtain correspondence between voxel[2] weight activations *and* associate feature weights on the image stimuli such that the relation between cognitive activity and task-stimuli (and vice-versa) is preserved. We speculate, as in Hardoon et al. (2007), that such a learnt relationship may better help our understanding of the neurological responses to a cognitive task and similarly to understand the image features that relate to these responses. The latter could assist in devising better fMRI experiments.

In this paper we present a novel and straightforward approach towards decomposing the tensor kernel SVM weight vector without accessing the feature spaces. We show that the performance of the method is statistically indistinguishable from the original tensor method, demonstrating that the decomposition does not impair the classification accuracy. Furthermore, we show that the decomposed weights can also be used as single source classifiers as well as for performing content based information retrieval. The neurological interpretation of the resulting weight maps (Mourão-Miranda et al. 2005, 2006; Hardoon et al. 2007) is outside the scope of this paper and will be addressed in a future study.

The paper is laid out as follows; In Sect. 2 we discuss the nomenclature used throughout the paper and briefly review the SVM formulation with a tensor kernel. Our main and novel results are given in Sect. 3, where we show how the decomposition of the weight vector derived from a SVM with a tensor kernel can be computed without accessing the feature spaces. Section 4 focuses on the issue of how to select an appropriate subspace for the decomposition while in Sect. 5 we elaborate on our experiments with the paired fRMI and image stimuli data. The paper is concluded with a discussion in Sect. 6.

2 Nomenclature & SVM formulation

We begin by introducing in Table 1 the general nomenclature used throughout the paper. Furthermore, we consider samples from a pair of random vectors (i.i.d. assumptions hold) of the form $(\mathbf{x}_i, \mathbf{y}_i)$ each with zero mean (i.e. centered) where $i = 1, \ldots, m$.

We first quote from Cristianini and Shawe-Taylor (2000) the general dual SVM optimisation as

$$\begin{aligned} &\max_{\boldsymbol{\alpha}} W(\boldsymbol{\alpha}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j c_i c_j \kappa(\mathbf{x}_i, \mathbf{x}_j) \\ &\text{subject to} \quad \sum_{i=1}^{m} \alpha_i c_i = 0 \quad \text{and} \quad \alpha_i \geq 0, \quad i = 1, \ldots, m, \end{aligned} \tag{1}$$

where we use c_i to represent the label of the data rather than the conventional y_i (this is because we use $\mathbf{y}_i$ to represent the paired sample of $\mathbf{x}_i$). The resulting decision function is

$$f(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i c_i \kappa(\mathbf{x}_i, \mathbf{x}). \tag{2}$$

[1] We limit ourselves to tensor kernels constructed of only two sources.

[2] A voxel is a pixel representing the smallest three-dimensional point volume referenced in an fMRI image of the brain. It is usually approximately 3 mm × 3 mm.

Table 1 General nomenclature used throughout the paper

s	lower case represents a scalar
$\mathbf{v}$	bold face represents a vector
M	upper case represents a matrix
κ	represents a kernel function
ϕ	projection operator into feature space
$\circ$	is the tensor product
U	a matrix whose columns are the vectors $\mathbf{u}$
m	denotes the number of samples
c	class label

It is easily shown how a vector space tensor product can be turned into an inner product space

$$\langle \phi_x \circ \phi_y, \psi_x \circ \psi_y \rangle_{H_x \circ H_y} = \langle \phi_x, \psi_x \rangle_{H_x} \langle \phi_y, \psi_y \rangle_{H_y}$$

for all $\phi_x, \psi_x \in H_x$ and $\phi_y, \psi_y \in H_y$ where H_x, H_y are two Hilbert spaces. Therefore the tensor product between $\mathbf{x}_i, \mathbf{y}_i$ can be represented as point-wise dot product kernel between the two respective kernels $\kappa(\mathbf{x}_i \circ \mathbf{y}_i, \mathbf{x}_j \circ \mathbf{y}_j) = \kappa_x(\mathbf{x}_i, \mathbf{x}_j)\kappa_y(\mathbf{y}_i, \mathbf{y}_j)$ and hence, the change in the SVM optimisation in (1) is only in the kernel used. Let $\hat{\kappa}$ be the tensor kernel matrix, so that

$$\hat{\kappa}_{ij} = \kappa_x(\mathbf{x}_i, \mathbf{x}_j)\kappa_y(\mathbf{y}_i, \mathbf{y}_j).$$

Detailed description of tensor products and their operations in Hilbert spaces are given in Szedmak et al. (2005), Pulmannová (2004) and therefore omitted here. Throughout the paper we assume that the resulting eigenvalues from the symmetric eigenproblem $A\mathbf{x} = \lambda\mathbf{x}$ are ordered such that $\lambda_x \geq \lambda_y \geq \cdots \geq \lambda_\ell$.

3 Tensor decomposition

We now give the main focus and novel contribution of the paper. The goal is to decompose the weight matrix W given by a dual representation $W = \sum_i^m \alpha_i c_i \phi_x(\mathbf{x}_i) \circ \phi_y(\mathbf{y}_i)$ without accessing the feature space. Given the paired samples $\mathbf{x}, \mathbf{y}$ the decision function in (2) becomes

$$f(\mathbf{x}, \mathbf{y}) = W \circ \phi_x(\mathbf{x})\phi_y(\mathbf{y})' = \sum_{i=1}^{m} \alpha_i c_i \kappa_x(\mathbf{x}_i, \mathbf{x})\kappa_y(\mathbf{y}_i, \mathbf{y}),$$

where we are able to express

$$W = \sum_{i=1}^{m} \alpha_i c_i \phi_x(\mathbf{x}_i)\phi_y(\mathbf{y}_i)'.$$

We want to decompose the weight matrix into a sum of tensor products of corresponding weight components for H_x and H_y

$$W \approx W^T = \sum_{t=1}^{T} \mathbf{w}_x^t \mathbf{w}_y^{t\prime}, \tag{3}$$

where for $t = 1, \ldots, T$ (we address the selection of T, the number of projections used, later in the paper) we have

$$\mathbf{w}_x^t \in \text{span}\,(\phi_x(\mathbf{x}_i), 1 \le i \le m) \subseteq H_x,$$
$$\mathbf{w}_y^t \in \text{span}\left(\phi_y(\mathbf{y}_i), 1 \le i \le m\right) \subseteq H_y,$$

so that $\mathbf{w}_x^t = \sum_{i=1}^m \beta_i^t \phi_x(\mathbf{x}_i)$ and $\mathbf{w}_y^t = \sum_{i=1}^m \gamma_i^t \phi_y(\mathbf{y}_i)$ where $\boldsymbol{\beta}^t$, $\boldsymbol{\gamma}^t$ are the dual variables of $\mathbf{w}_x^t, \mathbf{w}_y^t$. We compute

$$\begin{aligned} WW' &= \sum_{i=1}^m \alpha_i c_i \phi_x(\mathbf{x}_i)\phi_y(\mathbf{y}_i)' \sum_{j=1}^m \alpha_j c_j \phi_y(\mathbf{y}_j)\phi_x(\mathbf{x}_j)' \\ &= \sum_{i,j}^m \alpha_i \alpha_j c_i c_j \phi_x(\mathbf{x}_i)\kappa_y(\mathbf{y}_i, \mathbf{y}_j)\phi_x(\mathbf{x}_j)' \\ &= \sum_{i,j}^m \alpha_i \alpha_j c_i c_j \kappa_y(\mathbf{y}_i, \mathbf{y}_j)\phi_x(\mathbf{x}_i)\phi_x(\mathbf{x}_j)' \end{aligned} \tag{4}$$

and are able to express

$$K^y = \left(\kappa_y(\mathbf{y}_i, \mathbf{y}_j)\right)_{i,j=1}^m = \sum_{k=1}^K \lambda_k \mathbf{u}^k \mathbf{u}^{k\prime} = U \Lambda U', \tag{5}$$

where $U = (\mathbf{u}_1, \ldots, \mathbf{u}_K)$ by performing an eigenvalue decomposition of the kernel matrix K^y with entries $K_{ij}^y = \kappa_y(\mathbf{y}_i, \mathbf{y}_j)$. Substituting back into (4) gives

$$WW' = \sum_k^K \lambda_k \sum_{i,j}^m \alpha_i \alpha_j c_i c_j \mathbf{u}_i^k \mathbf{u}_j^{k\prime} \phi_x(\mathbf{x}_i)\phi_x(\mathbf{x}_j)'.$$

Letting $\mathbf{h}_k = \sum_{i=1}^m \alpha_i c_i \mathbf{u}_i^k \phi_x(\mathbf{x}_i)$ we have

$$WW' = \sum_k^K \lambda_k \mathbf{h}_k \mathbf{h}_k' = HH',$$

where $H = (\sqrt{\lambda_1}\mathbf{h}_1, \ldots, \sqrt{\lambda_K}\mathbf{h}_K)$. Note that $H \neq W$ but is a low dimensional representation of W (similar to performing a PCA). We would like to find the singular value decomposition of $H = V \Upsilon Z'$. Consider for $A = \text{diag}(\boldsymbol{\alpha})$ and $C = \text{diag}(\mathbf{c})$ we have

$$\begin{aligned} \left[H'H\right]_{k\ell} &= \sqrt{\lambda_k \lambda_\ell}\mathbf{h}_k' \mathbf{h}_\ell \\ &= \sqrt{\lambda_k \lambda_\ell} \sum_{ij} \alpha_i \alpha_j c_i c_j \mathbf{u}_i^k \mathbf{u}_j^\ell \kappa_x(\mathbf{x}_i, \mathbf{x}_j) \\ &= \left[\left(CAU\Lambda^{\frac{1}{2}}\right)' K^x \left(CAU\Lambda^{\frac{1}{2}}\right)\right]_{k\ell}, \end{aligned}$$

which is computable without accessing the feature space. Performing an eigenvalue decomposition on $H'H$ we have

$$H'H = Z\Upsilon V' V \Upsilon Z' = Z\Upsilon^2 Z' \tag{6}$$

with Υ a matrix with υ_t on the diagonal truncated after the J'th eigenvalue, which gives the dual representation of

$$\mathbf{v}_t = \frac{1}{\upsilon_t} H\mathbf{z}_t, \quad t = 1, \ldots, T,$$

and since

$$H'H\mathbf{z}_t = \upsilon_t^2 \mathbf{z}_t$$

we are able to verify that

$$WW'\mathbf{v}_t = HH'\mathbf{v}_t = \frac{1}{\upsilon_t} HH'H\mathbf{z}_t = \upsilon_t H\mathbf{z}_t = \upsilon_t^2 \mathbf{v}_t.$$

We are now able to express W as

$$W = IW = \left(\sum_{t=1}^{m} \mathbf{v}_t \mathbf{v}_t'\right) W = \sum_{t=1}^{m} \mathbf{v}_t \left(\mathbf{v}_t' W\right) = \sum_{t=1}^{m} \mathbf{v}_t \left(W'\mathbf{v}_t\right)'.$$

Restricting to the first T singular vectors allows us to express

$$W \approx W^T = \sum_{t=1}^{T} \mathbf{v}_t \left(W'\mathbf{v}_t\right)',$$

and from (3) we are able to express $\mathbf{w}_x^t = \mathbf{v}_t$ and $\mathbf{w}_y^t = W'\mathbf{v}_t$, which in turn results in

$$\begin{aligned}
\mathbf{w}_x^t = \mathbf{v}_t &= \frac{1}{\upsilon_t} H\mathbf{z}_t \\
&= \frac{1}{\upsilon_t} \sum_{k=1}^{T} \sqrt{\lambda_k} \sum_{i=1}^{m} \alpha_i c_i u_i^k \phi_x(\mathbf{x}_i) \mathbf{z}_k^t \\
&= \sum_{i=1}^{m} \left(\frac{1}{\upsilon_t} \alpha_i c_i \sum_{k=1}^{T} \sqrt{\lambda_k} \mathbf{z}_k^t u_i^k \right) \phi_x(\mathbf{x}_i) \\
&= \sum_{i=1}^{m} \beta_i^t \phi_x(\mathbf{x}_i),
\end{aligned}$$

where $\beta_i^t = \frac{1}{\upsilon_t} \alpha_i c_i \sum_{k=1}^{T} \sqrt{\lambda_k} \mathbf{z}_k^t u_i^k$. We can now also express

$$\begin{aligned}
\mathbf{w}_y^t = W'\mathbf{v}_t &= \frac{1}{\upsilon_t} W'H\mathbf{z}_t \\
&= W' \sum_{i=1}^{m} \beta_i^t \phi_x(\mathbf{x}_i) \\
&= \sum_{i=1}^{m} \alpha_i c_i \phi_y(\mathbf{y}_i) \sum_{j=1}^{m} \kappa(\mathbf{x}_i, \mathbf{x}_j) \beta_j^t
\end{aligned}$$

$$= \sum_{i=1}^{m} \left(\sum_{j=1}^{m} \alpha_i c_i \beta_j^t \kappa_x(\mathbf{x}_i, \mathbf{x}_j) \right) \phi_y(\mathbf{y}_i)$$

$$= \sum_{i=1}^{m} \gamma_i^t \phi_y(\mathbf{y}_i),$$

where $\gamma_i^t = \sum_{j=1}^{m} \alpha_i c_i \beta_j^t \kappa_x(\mathbf{x}_i, \mathbf{x}_j)$ are the dual variables of $\mathbf{w}_y^t$. We are therefore now able to decompose W into W_x, W_y without accessing the feature space giving us the desired result.

We continue to construct a new feature representation for a sample $\mathbf{x}$ as

$$\hat{\phi}_x(\mathbf{x}) = \left[\sum_{i=1}^{m} \kappa_x(\mathbf{x}_i, \mathbf{x}) \beta_i^t \right]_{t=1}^{T} \tag{7}$$

and similarly are able to construct the new feature representation for a sample $\mathbf{y}$ from the second view as

$$\hat{\phi}_y(\mathbf{y}) = \left[\sum_{i=1}^{m} \kappa_y(\mathbf{y}_i, \mathbf{y}) \gamma_i^t \right]_{t=1}^{T}. \tag{8}$$

Proposition 1 *With the computation described above we have that*

$$f(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{m} \alpha_i c_i \kappa_x(\mathbf{x}_i, \mathbf{x}) \kappa_y(\mathbf{y}_i, \mathbf{y})$$

$$\approx W^T (\phi_x(\mathbf{x}) \circ \phi_y(\mathbf{y}))$$

$$= \hat{\phi}_x(\mathbf{x})' \hat{\phi}_y(\mathbf{y})$$

with equality if T is chosen to be m.

We refer to the above tensor decomposition procedure as TD.

4 Subspace selection

We observe that the decomposition in Sect. 3 leaves open the question of selecting the number of eigenvectors to construct the subspaces for the decomposition. This corresponds to the selection of K and T in (5) and (6) respectively. We observe that T in (3) is determined by the eigenvalue decomposition in (6). Therefore, by ensuring that $T \leq \min(\text{rank}(K^y), \text{rank}(K^x))$ we can select $T = K$ in (6) unless there is a non-trivial intersection between the span of $\{\mathbf{u}_1, \ldots, \mathbf{u}_K\}$ and K^y.

We propose the following two approaches in setting the number of projections K. The first approach is by using the PCA bound proposed in Shawe-Taylor et al. (2005), which motivates selecting K projections that no longer improve the bound. We quote the bound in Theorem 1.

Theorem 1 (Shawe-Taylor et al. 2005) *If we perform PCA in the feature space defined by a kernel $\kappa(\mathbf{x}, \mathbf{z})$ then with probability great than $1 - \delta$, for any $1 \leq k \leq m$, if we project new*

data onto the space U_k, the expected squared residual is bounded by

$$\mathbb{E}\left[\|P_{U_k}^{\perp}(\boldsymbol{\phi}(\mathbf{x}))\|^2\right] \leq \min_{1\leq t\leq k}\left[\frac{1}{m}\lambda^{>t}(S) + \frac{1+\sqrt{t}}{\sqrt{m}}\sqrt{\frac{2}{m}\sum_{i=1}^{m}\kappa(\mathbf{x}_i,\mathbf{x}_i)}\right] + R^2\sqrt{\frac{18}{m}\ln\left(\frac{2m}{\delta}\right)},$$

where the support of the distribution is in a ball of radius R in the feature space and $\lambda^{>t}(S) = \sum_{i=t+1}^{m}\lambda_i$ is the sum of the eigenvalues greater than t computed from the training data in the feature space.

We observe that the value of the bound, for a kernel and fixed δ, will be shifted by a constant for all values of K normalised. Therefore, we use a simplified version of the bound

$$\tilde{\mathbb{E}} \leq \min_{1\leq t\leq k}\left[\frac{1}{m}\lambda^{>t}(S) + \frac{2(1+\sqrt{t})}{\sqrt{m}}\right],$$

which is sufficient to indicate when adding more eigenvalues will not improve the bound.

The second approach is to take the maximum number of eigenvectors corresponding to non-zero eigenvalues. We refer to this method as 'Max' in the experiments.

5 Experiments

fMRI data[3] was acquired from 16 right handed healthy US college male students aged 20–25 who, according to a self report, did not have any history of neurological or psychiatric illness. The subjects viewed image stimuli of three different active conditions: viewing unpleasant (dermatologic diseases), neutral (people), pleasant images (female models in swimsuits and lingerie), and a control condition (fixation). In these experiments only unpleasant and pleasant image categories are used.

The image-stimuli were presented in a block fashion and consisted of 42 images per category. During the experiment, there were 6 blocks of each active condition (each consisting of 7 image volumes) alternating with control blocks (fixation) of 7 images volumes. Similarly to the work in Hardoon et al. (2007) we associate pleasant with positive and unpleasant with negative and represent the image stimuli using the Scale Invariant Feature Transformation (SIFT) (Lowe 1999). Furthermore, we apply conventional pre-processing to the fMRI data. A detailed description of the fMRI pre-processing procedure and image-stimuli representation is given in Hardoon et al. (2007).

We run the experiments in a leave-subject-out fashion where 15 subjects are combined for training and a single subject is withheld for testing. This gave a sum total of $42\times 2\times 15 = 1260$ training and $42\times 2 = 84$ testing fMRI volumes and paired image stimuli. The analysis was repeated 16 times using linear kernels. We use the LIBSVM 2.85[4] (Fan et al. 2005) package for our SVM computation using default parameters and run all our experimentation on a 3 GHz 2×Xeon X5450 Quad Cores with 32 Gb RAM running on Centos 4.5.

We find that the bound in Theorem 1 indicates that no improvement can be achieved after the first 50 eigenvalues. The value of the simplified bound is plotted in Fig. 1.

[3]The data was acquired in a study by Mourão-Miranda et al. (2006).

[4]http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

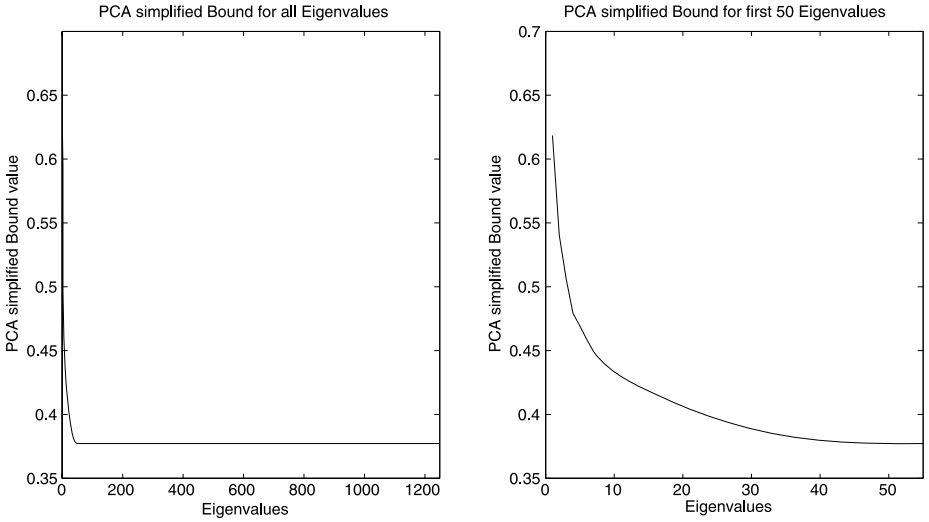


Fig. 1 We plot the simplified PCA bound values for the different number of eigenvalues decomposition of K^y. We are able to observe that roughly after 50 eigenvalues we are no longer able to improve on the bound, indicating this to be a sufficient number of eigenvectors (corresponding to the largest 50 eigenvalues) for our computation

In the following experiments we propose to decompose the tensor kernel into the dual variables β^t, γ^t in order to construct a new feature representation using (7) and (8) for the two sources respectively. This procedure enables us to project the data, from individual sources, into the common semantic space learnt by the tensor kernel SVM. In other words, we can construct a new kernel $\hat{K}_x = \langle \hat{\phi}_x(\mathbf{x}), \hat{\phi}_x(\mathbf{x}) \rangle$ which can be trained as usual using an SVM. We refer to our proposed procedure as the Decomposed Tensor kernel SVM (DTS) and proceed to demonstrate, in the following sections, that the DTS is able to maintain good results in comparison to the tensor kernel SVM. Whereas now the TD procedure allows us to compute individual weight vectors which have the potential, in our case, to elucidate the relationship between the fMRI activation and the image-stimuli features.

5.1 Results

5.1.1 Paired data

Initially, we aim to demonstrate that DTS does not lose, as a result of the decomposition, any discriminability from the original tensor SVM. We show this by combining the decomposed features back into a tensor product kernel and training, as well as testing, an SVM. To avoid any ambiguity we refer to this as DTS. Our baseline is the tensor kernel SVM trained on the paired fMRI and image-stimuli samples, for brevity we refer to this as the tensor SVM. We also consider an SVM in which the two feature vectors were concatenated into one high dimensional vector. We refer to this as SVM concatenated. Furthermore we compare to kernel CCA (kCCA) (Bach and Jordan 2002; Hardoon et al. 2004) where the learnt projections, of the two views, are used to create a tensor kernel, on which an SVM is trained and tested.[5]

[5]We use all the learnt kCCA directions for projection into the common semantic feature space.

Table 2 Results on the leave-subject-out procedure across the 16 subjects. We compare DTS (with the two approaches to the subspace selection) to an SVM trained on the fMRI + Image Stimuli feature spaces concatenated, an SVM trained on the kCCA joint semantic space (where the learnt projections are used to create a tensor kernel) and an SVM using the tensor kernel

Subject	SVM concat.	kCCA SVM	Tensor SVM	DTS	
				Th. 1	Max
Sub 01	86.90	80.95	89.28	88.09	89.28
Sub 02	80.95	73.80	80.95	88.09	86.90
Sub 03	79.76	75.00	75.00	76.19	75.00
Sub 04	83.33	77.38	95.23	90.47	91.66
Sub 05	80.95	70.23	88.09	90.47	91.66
Sub 06	83.33	80.95	88.09	84.52	85.71
Sub 07	82.14	75.00	89.28	89.28	90.47
Sub 08	76.19	72.61	83.33	90.47	88.09
Sub 09	69.04	63.09	72.61	82.14	82.14
Sub 10	70.23	70.23	89.28	83.33	83.33
Sub 11	88.09	78.57	84.52	84.52	82.14
Sub 12	82.14	75.00	86.90	88.09	83.33
Sub 13	75.00	79.76	86.90	89.28	88.09
Sub 14	55.95	57.14	86.90	94.04	92.85
Sub 15	75.00	67.85	86.90	84.52	83.33
Sub 16	85.71	83.33	91.66	95.23	95.23
average	78.42 ± 8.15	73.80 ± 6.94	85.93 ± 5.75	87.42 ± 4.74	86.83 ± 5.15

The results of this comparison are given in Table 2 where we are able to observe that the tensor SVM significantly improves on the concatenation of the two feature vectors as well as the kCCA joint semantic space learnt in an unsupervised fashion. Furthermore, we compare to DTS using the two proposed approaches for subspace selections (described in Sect. 4), and are able to observe that the decomposition maintains the original quality of discrimination.

5.1.2 Single source data

In the previous comparison we used paired-samples for the training and testing. In this section we wish to test the quality of the decomposed subspace by comparing our proposed method to the following two baselines where only the fMRI samples are used for testing;

1. A vanilla SVM trained on the fMRI samples.
2. kCCA trained on the paired-data followed by an SVM trained only on the fMRI samples projected into the learnt kCCA semantic space.
3. The tensor SVM trained on the paired fMRI and image-stimuli samples but only tested using fMRI testing samples. Conventionally the tensor kernel SVM uses paired samples for testing, although for this case we assume the paired image-stimuli test kernel to be an all ones matrix (i.e. we test using only the fMRI test samples).

In Table 3 we compare DTS, using an SVM trained only on the decomposed fMRI source, to the three baselines detailed above. We are able to observe that our proposed

Table 3 Results on the leave-subject-out procedure across the 16 subjects. We compare the DTS (with the two approaches to the subspace selection) to a vanilla SVM, an SVM trained only on fMRI samples projected into the common kCCA space as well as to an SVM using the tensor kernel. The testing procedure across all methods only involved the fMRI testing samples

Subject	fMRI SVM	kCCA SVM	Tensor SVM	DTS	
				Th. 1	Max
Sub 01	86.90	83.33	71.42	80.95	83.33
Sub 02	80.95	72.61	84.52	86.90	91.66
Sub 03	79.76	82.14	66.67	76.19	78.57
Sub 04	83.33	78.57	84.52	86.90	85.71
Sub 05	80.95	70.23	77.38	78.57	76.19
Sub 06	83.33	84.52	72.61	85.71	84.52
Sub 07	82.14	77.38	69.04	76.19	77.38
Sub 08	76.19	71.42	65.47	71.42	70.23
Sub 09	69.04	66.66	57.14	67.85	65.47
Sub 10	70.23	69.04	64.28	75.00	71.42
Sub 11	88.09	82.14	77.38	79.76	79.76
Sub 12	82.14	77.38	72.61	82.14	80.95
Sub 13	75.00	76.19	73.80	88.09	89.28
Sub 14	55.95	55.95	63.09	64.28	63.09
Sub 15	75.00	70.23	72.61	72.61	76.19
Sub 16	85.71	80.95	68.04	83.33	83.33
mean	78.42	74.92	71.35	78.50	78.57
std	±8.15	±7.54	±7.38	±7.08	±8.03

method achieves a improvement on the tensor SVM and kCCA SVM accuracy, while obtaining similar results to the vanilla SVM. These results indicate that the tensor SVM is unable to perform well when tested only using a single source. Whereas our method is able to learn the joint semantic space using the two sources and successfully decompose it to its individual components that retain a high classification accuracy when tested independently.

It is interesting to observe that in the case of six subjects[6] the vanilla SVM trained and tested on a single source performs better than the methods using both sources. We speculate that in these cases the fMRI data is sufficient to discriminate between the tasks. Although for eight subjects, we find that exploiting information from both sources improves the discrimination. In particular, we focus on subject 14 where our proposed approach improved the accuracy by 8.33%. While this is not the only occurrence where such a large improvement is achieved, it is the only case where SVM obtains a near random discrimination of 55.95%. Furthermore, we are able to observe that kCCA does not obtain any improvement, suggesting that explicit task information is important in extracting information from two sources.

Based on the poor SVM performance, it was hypothesised that subject 14 did not have a significantly different cognitive response to image-stimuli of ‘female models in swimsuits’ and that of ‘dermatologic diseases’. However, following the improvement on Subject 14’s

[6]Subjects number {1, 7, 8, 9, 11, 13}.

accuracy by our proposed approach, we now speculate that the reduced SVM ability in discrimination is potentially due to an increased level of biological noise within Subject 14's scans. The improvement may be achieved due to noise being filtered out during the weight decomposition. The hypothesis of biological-noise hampering the classification will be addressed in a separate study.

Furthermore, it is interesting to observe that the results obtained by 'fMRI SVM' are the same as those listed under 'SVM concatenated' in Table 2. Further investigation had revealed that this is due to the weight vectors of the two methods[7] being identical. This observation will also be investigated in a future follow-up analysis.

5.2 Weight maps in voxel space

We continue with the visualisation of the computed weight maps in voxel space for the tensor decomposition and subsequent SVM methods. The computed 'SVM concatenated' weight vector (corresponding to the voxel space) was found to be equal to the weight vector for the vanilla SVM trained only using fMRI data and was therefore omitted.

The brain regions identified in Sects. 5.2.1 and 5.2.2 were previously shown[8] in the neuroimaging literature to be associated with responses to visual and emotional processing. The detailed analysis and interpretation of the brain regions is beyond the scope of the paper and will be addressed in a separate study.

5.2.1 TD weight maps

In this section we visualise the weight maps in voxel space for the Tensor Decomposition (TD) (as detailed in Sect. 3). In Fig. 2 we plot $\mathbf{w}^t$ corresponding to the largest $t = 1, \ldots, 10$. Whereas, in Fig. 3 we plot the corresponding weight matrix from the tensor decomposition (TD) summed over the largest 10 components $\bar{\mathbf{w}} = \sum_{t=1}^{10} \mathbf{w}^t$ and averaged across all 16 subjects.

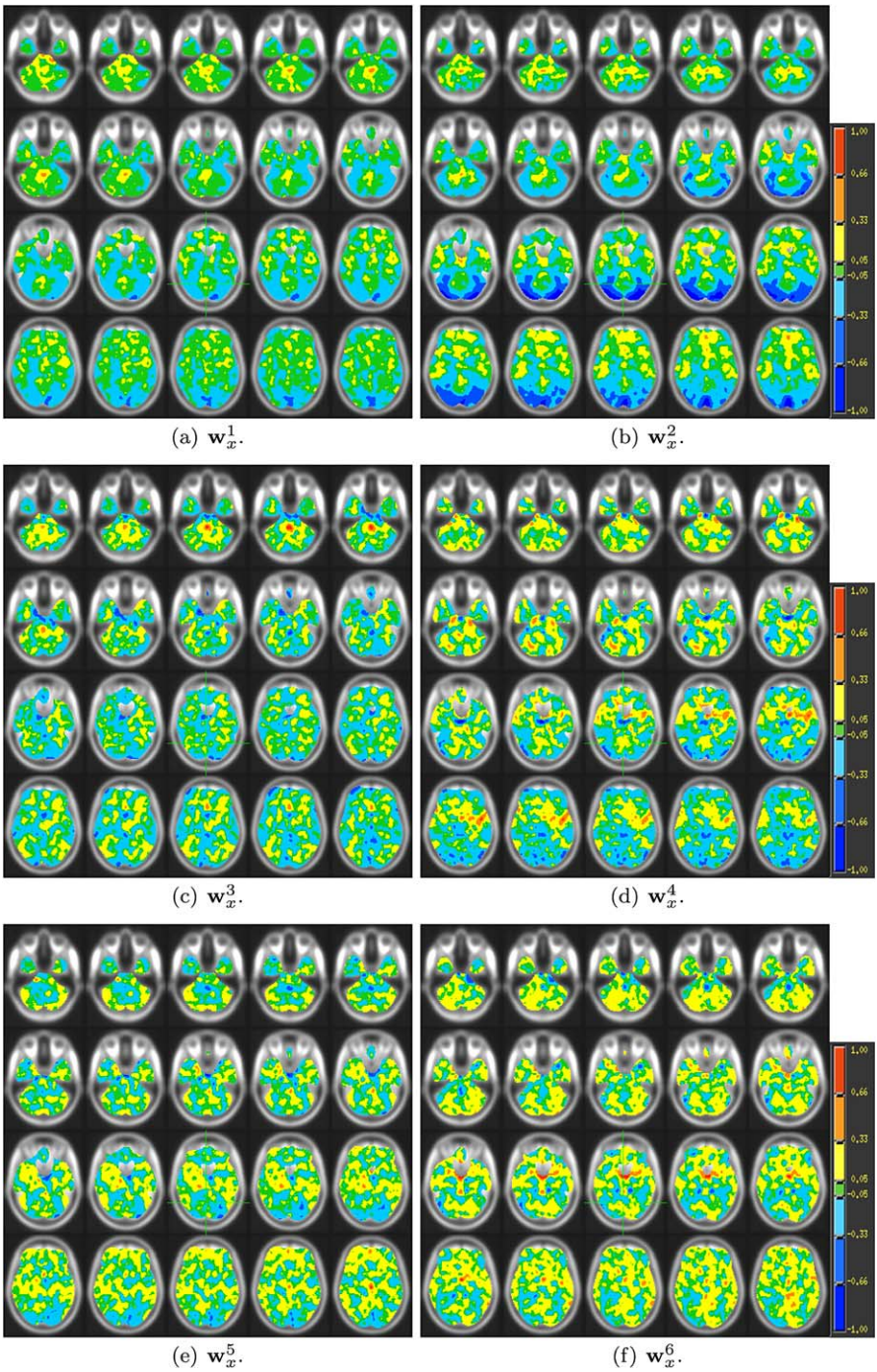
We are able to observe that the TD procedure allows us to compute individual weight vectors which have the potential to elucidate the relationship between the fMRI activation and the image-stimuli features. Furthermore, as the decomposition generates T directions we hypothesise that each of the resulting directions (map) correspond to particular neurological information that discriminate between the two tasks.

5.2.2 SVM weight maps

We continue to plot in Fig. 4 the weight maps in voxel space for SVM, kCCA-SVM, and DTS. The visualised weight vector $\mathbf{w}$ was computed as the average across the individual 16 weight vectors. We are able to observe that the weight map for both kCCA-SVM and DTS are a sparser than the SVM weight map. Furthermore we are able to observe that the DTS map emphasise regions not localised by kCCA-SVM.

[7] For the concatenated case, we are referring to the section of weight vector that corresponding to the fMRI features.

[8] Private communication.



(a) $\mathbf{w}_x^1$. (b) $\mathbf{w}_x^2$.

(c) $\mathbf{w}_x^3$. (d) $\mathbf{w}_x^4$.

(e) $\mathbf{w}_x^5$. (f) $\mathbf{w}_x^6$.

Fig. 2 (Color online) The unthresholded weight maps in voxel space showing the contrast between viewing pleasant vs. unpleasant for the top 10 decomposed tensor weights. We use the *blue* scale for negative (unpleasant) values and the *red* scale for the positive values (pleasant)

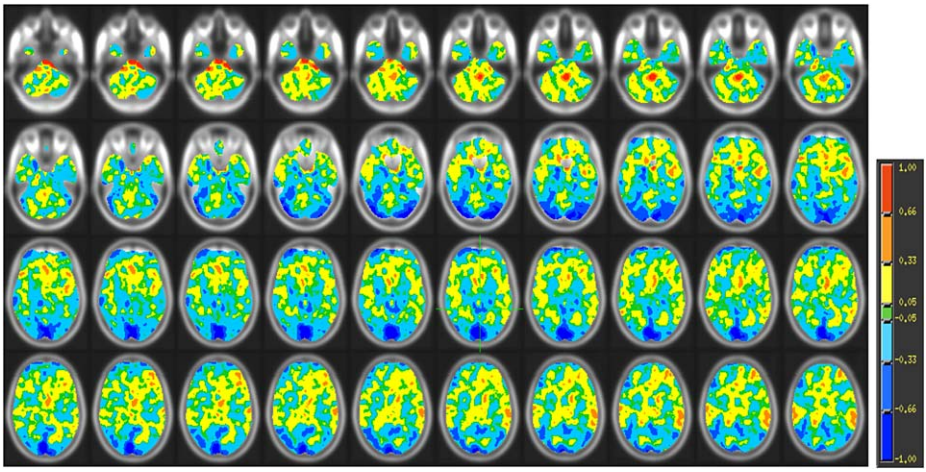


Fig. 3 (Color online) We plot, on axial slices, the unthresholded weight maps in voxel space showing the contrast between viewing pleasant vs. unpleasant for the summed top 10 decomposed tensor weights $\mathbf{w}_x = \sum_{t=1}^{10} \sum_{i=1}^{m} \beta_i^t \phi_x(\mathbf{x}_i)$. We use the *blue* scale for negative (unpleasant) values and the *red* scale for the positive values (pleasant)

5.3 Sparsity results

In the previous section we sequentially selected a consecutive number of features to represent the new features. In this section we use the Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani 1994) to select a sparse subset of the features to train and test (for simplicity we focus and compare to testing using only fMRI samples). We solve the LASSO problem using the framework proposed in Hardoon and Shawe-Taylor (2007).

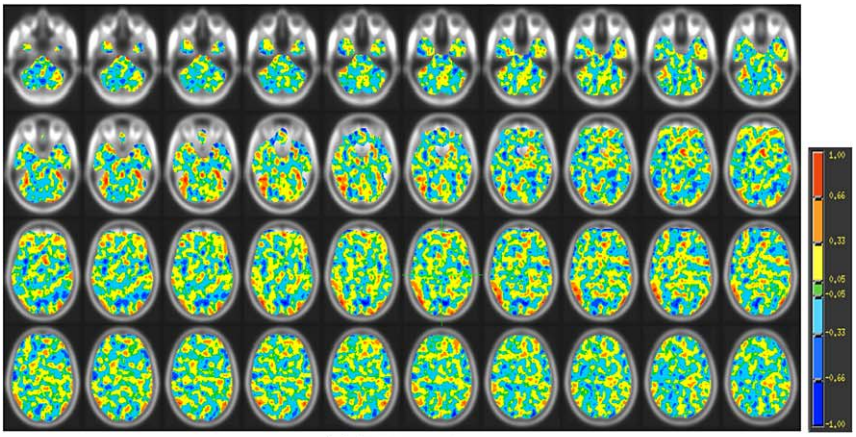
In Table 4 we are able to observe that the LASSO always chooses a number of features that is less than the maximum number of possible features (these correspond to the number of non-zero eigenvalues, i.e. rank of K^y). We note that even though the overall number of features used is identical to that indicated by the simplified bound, the LASSO does not necessarily choose consecutive features and this indeed improves the classification result.

5.4 Content retrieval

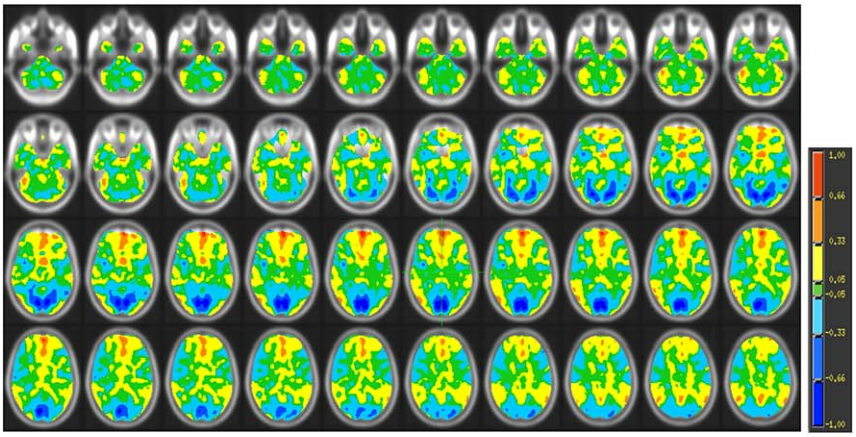
In this section we continue to verify the quality of the decomposed components in a content retrieval task. We compare DTS to kCCA for retrieval, where given a fMRI scan query i, we aim to retrieve a fMRI scan j (from the training corpus) with the same label as the query. Formally given as,

$$\max_{j \in m} \left\langle (\tilde{K}_i^x \beta), (K_j^x \beta) \right\rangle,$$

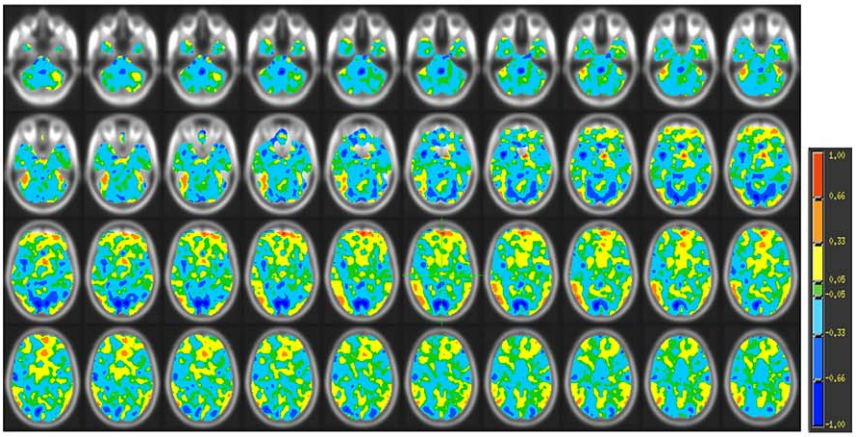
where $\tilde{K}_i^x$ is a vector product of a query scan i with the training data and β is either the DTS fMRI decomposed projections (as described in Sect. 3) or the kCCA fMRI projections. We give our results in Table 5 where we are able to observe the large improvement gained by DTS over kCCA. We believe these results indicate that the learnt projections by the tensor SVM, which are retained during the decomposition, provide a common semantic space that genuinely captures the underlying information of the tasks.



(a) SVM weight map.



(b) kCCA-SVM weight map.



(c) DTS weight map.

Fig. 4 (Color online) In the following subfigures we plot, on axial slices, the unthresholded weight maps in voxel space showing the contrast between viewing pleasant vs. unpleasant for; SVM in (**a**), kCCA-SVM in (**b**) and DTS in (**c**). We use the *blue* scale for negative (unpleasant) values and the *red* scale for the positive values (pleasant)

Table 4 We use a LASSO for selecting a sparse set of features for training and testing. Similarly to the previous experiments a leave-subject-out routine has been done across all 16 subjects and only fMRI testing samples have been used during the testing procedure

Subject	Accuracy	# features $\in 83$
Sub 01	82.14	29
Sub 02	90.48	29
Sub 03	84.52	32
Sub 04	85.71	36
Sub 05	79.76	38
Sub 06	83.33	83
Sub 07	78.57	83
Sub 08	72.62	83
Sub 09	71.43	83
Sub 10	79.76	43
Sub 11	82.14	31
Sub 12	82.14	38
Sub 13	88.10	38
Sub 14	65.48	83
Sub 15	79.76	83
Sub 16	82.14	43
average	80.51 ± 6.30	53.43

Table 5 Results on the leave-subject-out procedure across the 16 subjects for content retrieval. We compare the DTS (with the two approaches to the subspace selection) to kCCA. For both methods training was done using the paired data while the testing procedure only involved the fMRI testing samples

Subject	kCCA	DTS	
		Th. 1	Max
Sub 01	44.04	63.10	64.29
Sub 02	55.95	73.81	70.24
Sub 03	51.19	70.23	73.81
Sub 04	44.04	71.42	66.67
Sub 05	42.85	67.85	64.29
Sub 06	57.14	70.23	55.95
Sub 07	54.76	67.85	69.04
Sub 08	44.04	60.71	58.33
Sub 09	57.14	60.71	55.95
Sub 10	47.42	71.43	73.81
Sub 11	51.19	67.86	64.28
Sub 12	52.38	72.62	77.38
Sub 13	58.33	70.24	67.85
Sub 14	60.71	57.14	61.90
Sub 15	59.52	65.48	66.67
Sub 16	50.00	78.57	72.62
average	51.86 ± 6.10	68.08 ± 5.54	64.44 ± 6.38

6 Discussion

In this paper we address the issue of how to decompose a decision function learnt using the tensor kernel SVM with two sources, into their respective components such that we retain discriminability and obtain interpretability. The motivation and benefit for such a decomposition arrises from applications in the fields of Genetics, Neuroscience, Data mining, Psychometrics as well as others, where we wish to address more complex problems that require multi-source learning or analysis (e.g., Hardoon et al. 2007, 2009; Bickel et al. 2008) without sacrificing interpretability of the individual sources. We propose a novel approach for decomposing the resulting tensor weight into its two components without accessing the feature space.

We have demonstrated that DTS performs as well as the baseline approaches in three experiments of two-source and single-source classification while in a content retrieval task DTS outperforms KCCA based methods. This demonstrates that it is indeed possible to decompose the resulting tensor weight while retaining, and improving on, discriminability and, more importantly, *attaining* interpretability, the latter being important for practitioners.

In future studies we aim to address the application of this methodology to clinical studies as well as extending the theoretical understanding of the decomposition and its relationship to correlation analysis, as it is possible that the tensor space implicitly learns the correlation while discriminating between the tasks. Furthermore, we believe that the issue of sparsity and how it can further improve on discriminability is worth investigating as well as the extension of the tensor decomposition to more than two sources (Kolda and Sun 2008).

Acknowledgements David R. Hardoon[9] is supported by the EPSRC project Le Strum,[10] EP-D063612-1. The authors would like to thank Kristiaan Pelckmans for insightful discussions. This work was supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence,[11] IST-2007-216886. This publication only reflects the authors' views.

References

Anderson, D. R., Fite, K. V., Petrovich, N., & Hirsch, J. (2006). Cortical activation while watching video montage: An fMRI study. *Media Psychology*, *8*(1), 7–24.

Bach, F. R., & Jordan, M. I. (2002). Kernel independent component analysis. *Journal of Machine Learning Research*, *3*, 1–48.

Ben-Hur, A., & Noble, W. S. (2005). Kernel methods for predicting protein-protein interactions. *Bioinformatics*, *21*, i38–i46.

Bickel, S., Bogojeska, J., Lengauer, T., & Scheffer, T. (2008). Multi-task learning for HIV therapy screening. In *Proceedings of ICML*.

Carlson, T. A., Schrater, P., & He, S. (2003). Patterns of activity in the categorical representations of objects. *Journal of Cognitive Neuroscience*, *15*(5), 704–717.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge: Cambridge University Press.

Fan, R.-E., Chen, P.-H., & Lin, C.-J. (2005). Working set selection using the second order information for training SVM. *Journal of Machine Learning*, *6*, 1889–1918.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. P., Frith, C. D., & Frackowiak, R. S. J. (1995). Statistical parametric maps in functional imaging: a general linear approach. *Human Brain Mapping*, *2*(4), 189–210.

[9] http://www.davidroihardoon.com.

[10] http://www.lestrum.org.

[11] http://www.pascal-network.org.

Hardoon, D. R., & Shawe-Taylor, J. (2007). *Sparse canonical correlation analysis*. Technical report, University College London.

Hardoon, D. R., Szedmak, S., & Shawe-Taylor, J. (2004). Canonical correlation analysis: an overview with application to learning methods. *Neural Computation*, *16*(12), 2639–2664.

Hardoon, D. R., Mourão-Miranda, J., Brammer, M., & Shawe-Taylor, J. (2007). Unsupervised analysis of fMRI data using kernel canonical correlation. *NeuroImage*, *37*(4), 1250–1259.

Hardoon, D. R., Ettinger, U., Mourão-Miranda, J., Antonova, E., Collier, D., Kumari, V., Williams, S. C. R., & Brammer, M. (2009). Correlation based multivariate analysis of genetic influence on brain volume. *Neuroscience Letters*, *450*(3), 281–286.

Koelsch, S., Fritz, T., Yves, D., Cramon, V., Müller, K., & Friederici, A. D. (2006). Investigating emotion with music: An fMRI study. *Human Brain Mapping*, *27*(3), 239–250.

Kolda, T. G., & Sun, J. (2008). Scalable tensor decompositions for multi-aspect data mining. In *ICDM 2008: Proceedings of the 8th IEEE International Conference on Data Mining* (pp. 363–372), December 2008.

Kondor, R. I., & Lafferty, J. (2002). Diffusion kernels on graphs and other discrete input spaces. In *Proceedings of the Nineteenth International Conference on Machine Learning* (pp. 315–322). San Mateo: Morgan Kaufmann.

LaConte, S., Strother, S., Cherkassky, V., Anderson, J., & Hu, X. (2005). Support vector machines for temporal classification of block design fMRI data. *NeuroImage*, *26*(2), 317–329.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proceedings of the 7th IEEE International Conference on Computer Vision* (pp. 1150–1157), Kerkyra, Greece.

Martin, S., Roe, D., & Faulon, J.-L. (2005). Predicting protein-protein interactions using signature products. *Bioinformatics*, *21*(2), 218–226.

Mitchell, T. M., Hutchinson, R., Niculescu, R. S., Pereira, F., Wang, X., Just, M., & Newman, S. (2004). Learning to decode cognitive states from brain images. *Machine Learning*, *57*(1–2), 145–175.

Mourão-Miranda, J., Bokde, A. L. W., Born, C., Hampel, H., & Stetter, M. (2005). Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional MRI data. *NeuroImage*, *28*(4), 980–995.

Mourão-Miranda, J., Reynaud, E., McGlone, F., Calvert, G., & Brammer, M. (2006). The impact of temporal compression and space selection on SVM analysis of single-subject and multi-subject fMRI data. *NeuroImage*, *33*(4), 1055–1065.

O'Toole, A. J., Jiang, F., Abdi, H., Pénard, N., Dunlop, J. P., & Parent, M. A. (2007). Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. *Journal of Cognitive Neuroscience*, *19*(11), 1735–1752.

Pulmannová, S. (2004). Tensor products of Hilbert space effect algebras. *Reports on Mathematical Physics*, *53*(2), 301–316.

Qiu, J., & Noble, W. S. (2008). Predicting co-complexed protein pairs from heterogeneous data. *PLoS Computational Biology*, *4*(4), e1000054.

Shawe-Taylor, J., Williams, C. K. I., Cristianini, N., & Kandola, J. (2005). On the eigenspectrum of the Gram matrix and the generalization error of kernel-PCA. *IEEE Transactions on Information Theory*, *51*(7), 2510–2522.

Szedmak, S., Shawe-Taylor, J., & Parado-Hernandez, E. (2005). *Learning via linear operators: Maximum margin regression; multiclass and multiview learning at one-class complexity*. Technical report, University of Southampton.

Szedmak, S., De Bie, T., & Hardoon, D. R. (2007). A metamorphosis of canonical correlation analysis into multivariate maximum margin learning. In *Proceedings of the 15th European Symposium on Artificial Neural Networks (ESANN 2007)*, Bruges, April 2007.

Tibshirani, R. (1994). *Regression shrinkage and selection via the lasso*. Technical report, University of Toronto.

Weston, J., Baklr, G., Bousquet, O., Schölkopf, B., Mann, T., & Noble, W. S. (2007). Joint kernel maps. In G. Baklr, T. Hofmann, B. Scholkopf, A. J. Smola, B. Taskar, & S. V. N. Vishwanathan (Eds.), *Predicting structured data*. Cambridge: MIT Press.