A Unifying Framework for Vector-valued Manifold Regularization and Multi-view Learning

Hà Quang Minh MINH.HAQUANG@IIT.IT
Loris Bazzani LORIS.BAZZANI@IIT.IT
Vittorio Murino VITTORIO.MURINO@IIT.IT
Istituto Italiano di Tecnologia, Via Morego 30, Genova 16163, ITALY

Abstract

This paper presents a general vector-valued reproducing kernel Hilbert spaces (RKHS) formulation for the problem of learning an unknown functional dependency between a structured input space and a structured output space, in the Semi-Supervised Learning setting. Our formulation includes as special cases Vector-valued Manifold Regularization and Multi-view Learning, thus provides in particular a unifying framework linking these two important learning approaches. In the case of least square loss function, we provide a closed form solution with an efficient implementation. Numerical experiments on challenging multi-class categorization problems show that our multi-view learning formulation achieves results which are comparable with state of the art and are significantly better than single-view learning.

1. Introduction

Reproducing kernel Hilbert spaces (RKHS) and kernel-based methods have been by now established as among the most powerful paradigms in machine learning and statistics, with numerous practical applications, see e.g. (Schölkopf & Smola, 2002; Shawe-Taylor & Cristianini, 2004). While most of the literature on kernel methods so far has focused on scalar-valued functions, RKHS of vector-valued functions have received increasing research attention in machine learning recently, from both theoretical and practical perspectives, see e.g. (Micchelli & Pontil, 2005; Carmeli et al., 2006; Reisert & Burkhardt, 2007; Caponnetto et al., 2008; Brouard et al., 2011; Dinuzzo et al., 2011; Kadri et al., 2011; Minh & Sindhwani, 2011).

The goal of the present work is to fuse together two learning approaches for Semi-Supervised Learning in a general vector-valued RKHS framework, where the hypothesis spaces are RKHS of vector-valued functions.

The first approach for Semi-Supervised Learning that we consider is Manifold Regularization (Belkin et al., 2006), which attempts to learn the geometry of the input space from the given unlabeled data. Generalizations from the scalar case to the vector-valued setting include (Brouard et al., 2011), where a vector-valued version of the graph Laplacian L is used, and (Minh & Sindhwani, 2011), where L can be a general symmetric, positive operator, including the graph Laplacian. The vector-valued setting allows one to capture possible dependencies between output variables by the use of, for example, an output graph Laplacian.

The second approach for Semi-Supervised Learning that we consider is Multi-view Learning. In the multi-view approach (Brefeld et al., 2006; Sindhwani & Rosenberg, 2008; Rosenberg et al., 2009; Saffari et al., 2010), different hypothesis spaces are employed to construct target functions based on different aspects of the input data. These target functions not only need to agree with the labeled data, but also need to agree *with each other* on the unlabeled data. One scenario where this approach can be applied (Rosenberg et al., 2009) is when the input data can be decomposed as $x = (x^1, \dots, x^m)$, with each x^i representing one *view*. One can then construct a target function for each view and fuse them together to obtain the final solution.

The formulation we present in this paper gives a unified learning framework for the case the hypothesis spaces are vector-valued RKHS. Our formulation is general, encompassing many common algorithms as special cases, including both Vector-valued Manifold Regularization and Multi-view Learning. For the case

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28. Copyright 2013 by the author(s).

of least square loss function, we give a closed form solution which can be implemented efficiently. Our numerical experiments were performed using a special case of our framework, namely Vector-valued Multi-view Learning, with promising results on several particularly challenging multi-class categorization problems.

To the best of our knowledge, this work is the first attempt to present a unified general learning framework whose components have been only individually and partially covered in the literature.

Organization. We start by giving a review of vector-valued RKHS in Section 2. We state the general optimization problem we wish to solve in Section 3, along with various special cases, the Representer Theorem, and Proposition 1, which gives the explicit solution for the least square case. We then describe Vector-valued Multi-view Learning and its implementation in Section 4, with empirical experiments in Section 5. **Proofs for all mathematical results in the paper are given in the Supplementary Material.**

2. Vector-Valued RKHS

In this section, we give a brief review of RKHS of vector-valued functions[1], for more detail see e.g. (Carmeli et al., 2006; Micchelli & Pontil, 2005; Caponnetto et al., 2008; Minh & Sindhwani, 2011). In the following, denote by $\mathcal{X}$ a nonempty set, $\mathcal{Y}$ a real, separable Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{Y}}$, $\mathcal{L}(\mathcal{Y})$ the Banach space of bounded linear operators on $\mathcal{Y}$. Let $\mathcal{Y}^{\mathcal{X}}$ denote the vector space of all functions $f : \mathcal{X} \to \mathcal{Y}$. A function $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ is said to be an **operator-valued positive definite kernel** if for each pair $(x, z) \in \mathcal{X} \times \mathcal{X}$, $K(x, z)^* = K(z, x)$, and

$$\sum_{i,j=1}^{N} \langle y_i, K(x_i, x_j)y_j \rangle_{\mathcal{Y}} \geq 0 \qquad (1)$$

for every finite set of points $\{x_i\}_{i=1}^N$ in $\mathcal{X}$ and $\{y_i\}_{i=1}^N$ in $\mathcal{Y}$. Given such a K, there exists a unique $\mathcal{Y}$-valued RKHS $\mathcal{H}_K$ with reproducing kernel K, which is constructed as follows. For each $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, form a function $K_x y = K(., x)y \in \mathcal{Y}^{\mathcal{X}}$ defined by

$$(K_x y)(z) = K(z, x)y \quad \text{for all} \quad z \in X.$$

Consider the set $\mathcal{H}_0 = \text{span}\{K_x y \mid x \in \mathcal{X},\ y \in \mathcal{Y}\} \subset \mathcal{Y}^{\mathcal{X}}$. For $f = \sum_{i=1}^N K_{x_i} w_i$, $g = \sum_{i=1}^N K_{z_i} y_i \in \mathcal{H}_0$, we define the inner product

$$\langle\ f, g\ \rangle_{\mathcal{H}_K} = \sum_{i,j=1}^{N} \langle w_i, K(x_i, z_j) y_j \rangle_{\mathcal{Y}},$$

which makes $\mathcal{H}_0$ a pre-Hilbert space. Completing $\mathcal{H}_0$ by adding the limits of all Cauchy sequences gives the Hilbert space $\mathcal{H}_K$. The **reproducing property** is

$$\langle f(x), y \rangle_{\mathcal{Y}} = \langle f, K_x y \rangle_{\mathcal{H}_K} \quad \text{for all} \quad f \in \mathcal{H}_K. \qquad (2)$$

Sampling Operators. For each $x \in \mathcal{X}$, let $K_x : \mathcal{Y} \to \mathcal{H}_K$ be the operator with $K_x y$ defined as above, then

$$||K_x y||^2_{\mathcal{H}_K} = \langle K(x, x)y, y \rangle_{\mathcal{Y}} \leq ||K(x, x)||\ ||y||^2_{\mathcal{Y}},$$

which implies that

$$||K_x : \mathcal{Y} \to \mathcal{H}_K|| \leq \sqrt{||K(x, x)||},$$

so that K_x is a bounded operator. Let $K_x^* : \mathcal{H}_K \to \mathcal{Y}$ be the adjoint operator of K_x, then from (2), we have

$$f(x) = K_x^* f \qquad \text{for all} \quad x \in \mathcal{X}, f \in \mathcal{H}_K. \qquad (3)$$

From this we deduce that for all $x \in \mathcal{X}$ and all $f \in \mathcal{H}_K$,

$$||f(x)||_{\mathcal{Y}} \leq ||K_x^*||\ ||f||_{\mathcal{H}_K} \leq \sqrt{||K(x, x)||}\ ||f||_{\mathcal{H}_K},$$

that is the *sampling operator* $S_x : \mathcal{H}_K \to \mathcal{Y}$ defined by

$$S_x f = K_x^* f = f(x)$$

is bounded. Let $\mathbf{x} = (x_i)_{i=1}^l \in \mathcal{X}^l$, $l \in \mathbb{N}$. For the sampling operator $S_{\mathbf{x}} : \mathcal{H}_K \to \mathcal{Y}^l$ defined by $S_{\mathbf{x}}(f) = (f(x_i))_{i=1}^l$, for any $\mathbf{y} = (y_i)_{i=1}^l \in \mathcal{Y}^l$,

$$\begin{aligned}\langle S_{\mathbf{x}} f, \mathbf{y} \rangle_{\mathcal{Y}^l} &= \sum_{i=1}^{l} \langle f(x_i), y_i \rangle_{\mathcal{Y}} = \sum_{i=1}^{l} \langle K_{x_i}^* f, y_i \rangle_{\mathcal{H}_K} \\ &= \sum_{i=1}^{l} \langle f, K_{x_i} y_i \rangle_{\mathcal{H}_K} = \langle f, \sum_{i=1}^{l} K_{x_i} y_i \rangle_{\mathcal{H}_K}.\end{aligned}$$

Thus the adjoint operator $S_{\mathbf{x}}^* : \mathcal{Y}^l \to \mathcal{H}_K$ is given by

$$S_{\mathbf{x}}^* \mathbf{y} = S_{\mathbf{x}}^*(y_1, \ldots, y_l) = \sum_{i=1}^{l} K_{x_i} y_i, \quad \mathbf{y} \in \mathcal{Y}^l, \qquad (4)$$

and the operator $S_{\mathbf{x}}^* S_{\mathbf{x}} : \mathcal{H}_K \to \mathcal{H}_K$ is given by

$$S_{\mathbf{x}}^* S_{\mathbf{x}} f = \sum_{i=1}^{l} K_{x_i} f(x_i) = \sum_{i=1}^{l} K_{x_i} K_{x_i}^* f. \qquad (5)$$

Data-dependent Semi-norms. Let $(x_1, \ldots, x_{u+l}) \subset \mathcal{X}$. Let $M : \mathcal{Y}^{u+l} \to \mathcal{Y}^{u+l} \in \mathcal{L}(\mathcal{Y}^{u+l})$ be a symmetric, positive operator, that is $\langle y, My \rangle_{\mathcal{Y}^{u+l}} \geq 0$ for all

[1] Some authors, e.g. (Kadri et al., 2011) employ the terminology *function-valued*, which is equivalent to *vector-valued*: a function is a vector in a vector space of functions (e.g. a Hilbert space of functions), and an n-dimensional vector is a discrete function defined on n points.

$y \in \mathcal{Y}^{u+l}$. For $f \in \mathcal{H}_K$, let $\mathbf{f} = (f(x_1), \ldots, f(x_{u+l})) \in \mathcal{Y}^{u+l}$. The operator $M : \mathcal{Y}^{u+l} \to \mathcal{Y}^{u+l}$ can be expressed as an operator-valued matrix $M = (M_{ij})_{i,j=1}^{u+l}$ of size $(u+l) \times (u+l)$, with each $M_{ij} : \mathcal{Y} \to \mathcal{Y}$ being a linear operator, so that

$$(M\mathbf{f})_i = \sum_{j=1}^{u+l} M_{ij}\mathbf{f}_j = \sum_{j=1}^{u+l} M_{ij} f(x_j). \tag{6}$$

We can then define the following semi-norm for f, which depends on the x_i's:

$$\langle \mathbf{f}, M\mathbf{f} \rangle_{\mathcal{Y}^{u+l}} = \sum_{i,j=1}^{u+l} \langle f(x_i), M_{ij} f(x_j) \rangle_{\mathcal{Y}}. \tag{7}$$

This form of semi-norm was utilized in vector-valued manifold regularization (Minh & Sindhwani, 2011).

3. General Minimization Problem

In this section, we state the general minimization problem that we wish to solve, which includes Vector-valued Manifold Regularization and Multi-view Learning as special cases.

Let the input space be $\mathcal{X}$, an arbitrary non-empty set. Let $\mathcal{Y}$ be a separable Hilbert space, denoting the output space. Assume that there is an unknown probability measure ρ on $\mathcal{X} \times \mathcal{Y}$, and that we have access to a random training sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^{l} \cup \{x_i\}_{i=l+1}^{u+l}$ of l labeled and u unlabeled examples.

Let $\mathcal{W}$ be a separable Hilbert space. Let $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{W})$ be an operator-valued positive definite kernel and $\mathcal{H}_K$ its induced Reproducing Kernel Hilbert Space of $\mathcal{W}$-valued functions.

Let $M : \mathcal{W}^{u+l} \to \mathcal{W}^{u+l}$ be a symmetric, positive operator. For each $f \in \mathcal{H}_K$, let

$$\mathbf{f} = (f(x_1), \ldots, f(x_{u+l})) \in \mathcal{W}^{u+l}. \tag{8}$$

Let $V : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a convex loss function. Let $C : \mathcal{W} \to \mathcal{Y}$ be a bounded linear operator, with $C^* : \mathcal{Y} \to \mathcal{W}$ its adjoint operator.

The following is the general minimization problem that we wish to solve:

$$f_{\mathbf{z},\gamma} = \text{argmin}_{f \in \mathcal{H}_K} \frac{1}{l} \sum_{i=1}^{l} V(y_i, Cf(x_i)) + \gamma_A ||f||^2_{\mathcal{H}_K} + \gamma_I \langle \mathbf{f}, M\mathbf{f} \rangle_{\mathcal{W}^{u+l}}, \tag{9}$$

with regularization parameters $\gamma_A > 0$, $\gamma_I \geq 0$.

Let us give a general multi-view learning interpretation of our framework. If each input instance x has many views, then $f(x) \in \mathcal{W}$ represents the output values from all the views, constructed by their corresponding hypothesis spaces. These values are combined by the operator C to give the final output value in $\mathcal{Y}$, which is not necessarily the same as $\mathcal{W}$. In (9), the first term measures the error between the final output $Cf(x_i)$ for x_i with the given output y_i, $1 \leq i \leq l$. The second summand is the standard RKHS regularization term. The third summand, Multi-view Manifold Regularization, is a generalization of vector-valued Manifold Regularization in (Minh & Sindhwani, 2011) and Multi-view Point Cloud regularization in (Rosenberg et al., 2009): if there is only one view, then it is simply manifold regularization; if there are many views, then it consists of manifold regularization along each view, as well as consistency regularization across different views. We describe one concrete realization of this term in Section 4.2.

Remark 1. The framework is readily generalizable to the case the point evaluation functional $f(x)$ is replaced by a general bounded linear operator - we describe this in the Supplementary Material.

3.1. Representer Theorem

The minimization problem (9) is guaranteed to always have a unique global solution. The following is a natural generalization of the Representer Theorem in (Minh & Sindhwani, 2011).

Theorem 1. *The minimization problem (9) has a unique solution, given by $f_{\mathbf{z},\gamma} = \sum_{i=1}^{u+l} K_{x_i} a_i$ for some vectors $a_i \in \mathcal{W}$, $1 \leq i \leq u+l$.*

3.2. Least Square Case

For the case V is the least square loss function, we solve the following problem, which has an explicit solution:

$$f_{\mathbf{z},\gamma} = \text{argmin}_{f \in \mathcal{H}_K} \frac{1}{l} \sum_{i=1}^{l} ||y_i - Cf(x_i)||^2_{\mathcal{Y}} + \gamma_A ||f||^2_{\mathcal{H}_K} + \gamma_I \langle \mathbf{f}, M\mathbf{f} \rangle_{\mathcal{W}^{u+l}}. \tag{10}$$

The following is a generalization of Proposition 1 in (Minh & Sindhwani, 2011).

Proposition 1. *The minimization problem (10) has a unique solution $f_{\mathbf{z},\gamma} = \sum_{i=1}^{u+l} K_{x_i} a_i$, where the vectors $a_i \in \mathcal{W}$ are given by*

$$l\gamma_I \sum_{j,k=1}^{u+l} M_{ik} K(x_k, x_j) a_j + C^*C(\sum_{j=1}^{u+l} K(x_i, x_j) a_j) + l\gamma_A a_i = C^* y_i, \tag{11}$$

for $1 \leq i \leq l$, *and*

$$\gamma_I \sum_{j,k=1}^{u+l} M_{ik} K(x_k, x_j) a_j + \gamma_A a_i = 0, \qquad (12)$$

for $l+1 \leq i \leq u+l$.

3.3. Operator-valued Matrix Formulation

The system of equations (11) and (12) can be reformulated in matrix form, which is more readable and more convenient to implement efficiently.

Let $K[\mathbf{x}]$ denote the $(u+l) \times (u+l)$ operator-valued matrix whose (i,j) entry is $K(x_i, x_j)$. Let $J_l^{\mathcal{W},u+l} : \mathcal{W}^{u+l} \to \mathcal{W}^{u+l}$ denote the diagonal matrix whose first l entries on the main diagonal are the identity operator $I : \mathcal{W} \to \mathcal{W}$, with the rest being 0. Let $\mathbf{C}^*\mathbf{C} : \mathcal{W}^{u+l} \to \mathcal{W}^{u+l}$ be the $(u+l) \times (u+l)$ diagonal matrix, with each diagonal entry being $C^*C : \mathcal{W} \to \mathcal{W}$. Let $\mathbf{C}^* : \mathcal{Y}^{u+l} \to \mathcal{W}^{u+l}$ be the $(u+l) \times (u+l)$ diagonal matrix, with each diagonal entry being $C^* : \mathcal{Y} \to \mathcal{W}$. Then Proposition 1 is equivalent to

Proposition 2.

$$(\mathbf{C}^*\mathbf{C}J_l^{\mathcal{W},u+l}K[\mathbf{x}] + l\gamma_I MK[\mathbf{x}] + l\gamma_A I)\mathbf{a} = \mathbf{C}^*\mathbf{y}, \quad (13)$$

where $\mathbf{a} = (a_1, \ldots, a_{u+l})$, $\mathbf{y} = (y_1, \ldots, y_{u+l})$ *are considered as column vectors in* $\mathcal{W}^{u+l}$ *and* $\mathcal{Y}^{u+l}$, *respectively, and* $y_{l+1} = \cdots = y_{u+l} = 0$.

3.4. Special Cases

Vector-valued Regularized Least Squares. If $\mathbf{C}^*\mathbf{C} = I : \mathcal{W}^{u+l} \to \mathcal{W}^{u+l}$, then (13) reduces to

$$(J_l^{\mathcal{W},u+l}K[\mathbf{x}] + l\gamma_I MK[\mathbf{x}] + l\gamma_A I)\mathbf{a} = \mathbf{C}^*\mathbf{y}. \quad (14)$$

If $u = 0$, $\gamma_I = 0$, and $\gamma_A = \gamma$, then we have

$$(K[\mathbf{x}] + l\gamma I)\mathbf{a} = \mathbf{C}^*\mathbf{y}. \quad (15)$$

One particular case for this scenario is when $\mathcal{W} = \mathcal{Y}$ and $C : \mathcal{Y} \to \mathcal{Y}$ is a unitary operator, that is $C^*C = CC^* = I$. If $\mathcal{Y} = \mathbb{R}^n$ and $C : \mathbb{R}^n \to \mathbb{R}^n$ is real, then C is an orthogonal matrix. If $C = I$, then we recover the Regularized Least Squares algorithm.

Vector-valued Manifold Regularization. Let $\mathcal{W} = \mathcal{Y}$ and $C = I$. Then we obtain the minimization problem for vector-valued Manifold Regularization (Minh & Sindhwani, 2011):

$$f_{\mathbf{z},\gamma} = \text{argmin}_{f \in \mathcal{H}_K} \frac{1}{l} \sum_{i=1}^{l} V(y_i, f(x_i)) + \gamma_A ||f||^2_{\mathcal{H}_K} + \gamma_I \langle \mathbf{f}, M\mathbf{f} \rangle_{\mathcal{W}^{u+l}}. \quad (16)$$

Scalar Multi-view Learning. In this section, we show that the scalar multi-view learning formulation of (Sindhwani & Rosenberg, 2008; Rosenberg et al., 2009) can be cast as a special case of our framework. Let $\mathcal{Y} = \mathbb{R}$ and $k^1, \ldots, k^m$ be real-valued positive definite kernels on $\mathcal{X} \times \mathcal{X}$, with corresponding RKHS $\mathcal{H}_{k^i}$ of functions $f^i : \mathcal{X} \to \mathbb{R}$, with each $\mathcal{H}_{k^i}$ representing one view. Let $f = (f^1, \ldots, f^m)$, with $f^i \in \mathcal{H}_{k^i}$. Let $\mathbf{c} = (c_1, \ldots, c_m) \in \mathbb{R}^m$ be a fixed weight vector. In the notation of (Rosenberg et al., 2009), let $\mathbf{f} = (f^1(x_1), \ldots, f^1(x_{u+l}), \ldots, f^m(x_1), \ldots, f^m(x_{u+l}))$ and $M \in \mathbb{R}^{m(u+l) \times m(u+l)}$ be positive semidefinite. The objective of Multi-view Point Cloud Regularization (formula (4) in (Rosenberg et al., 2009)) is

$$\text{argmin}_{\varphi : \varphi(x) = \langle \mathbf{c}, f(x) \rangle} \frac{1}{l} \sum_{i=1}^{l} V(y_i, \varphi(x_i)) + \sum_{i=1}^{m} \gamma_i ||f^i||^2_{k^i} + \gamma \langle \mathbf{f}, M\mathbf{f} \rangle_{\mathbb{R}^{m(u+l)}}, \quad (17)$$

for some convex loss function V, with $\gamma_i > 0$, $i = 1, \ldots, m$, and $\gamma \geq 0$. Problem (17) admits a natural formulation in vector-valued RKHS. Let

$$K = \text{diag}(\frac{1}{\gamma_1}, \ldots, \frac{1}{\gamma_m}) * \text{diag}(k^1, \ldots, k^m) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{m \times m}, \quad (18)$$

then $f = (f^1, \ldots, f^m) \in \mathcal{H}_K : \mathcal{X} \to \mathbb{R}^m$, with

$$||f||^2_{\mathcal{H}_K} = \sum_{i=1}^{m} \gamma_i ||f^i||^2_{k^i}. \quad (19)$$

By the reproducing property, we have

$$\langle \mathbf{c}, f(x) \rangle_{\mathbb{R}^m} = \langle f, K_x \mathbf{c} \rangle_{\mathcal{H}_K}. \quad (20)$$

We can now recast (17) into

$$f_{\mathbf{z},\gamma} = \text{argmin}_{f \in \mathcal{H}_K} \frac{1}{l} \sum_{i=1}^{l} V(y_i, \langle \mathbf{c}, f(x) \rangle_{\mathbb{R}^m}) + ||f||^2_{\mathcal{H}_K} + \gamma \langle \mathbf{f}, M\mathbf{f} \rangle_{\mathbb{R}^{m(u+l)}}. \quad (21)$$

This is a special case of (9), with $\mathcal{W} = \mathbb{R}^m$, $\mathcal{Y} = \mathbb{R}$, and $C : \mathbb{R}^m \to \mathbb{R}$ given by

$$Cf(x) = \langle \mathbf{c}, f(x) \rangle_{\mathbb{R}^m} = c_1 f^1(x) + \cdots + c_m f^m(x). \quad (22)$$

The vector-valued formulation of scalar multi-view learning has the following advantages:

(i) The kernel K is diagonal matrix-valued and is obviously positive definite. In contrast, it is nontrivial to prove that the multi-view kernel of (Rosenberg et al., 2009) is positive definite.

(ii) The kernel K is independent of the c_i's, unlike the multi-view kernel of (Rosenberg et al., 2009), which needs to be recomputed for each different set c_i's.

(iii) One can recover all the component functions f^i's using K. In contrast, in (Sindhwani & Rosenberg, 2008), it is shown how one can recover the f^i's only when $m = 2$, but not in the general case.

4. Vector-valued Multi-view Learning

Another special case of our formulation, which is the focus of subsequent sections, is vector-valued multi-view learning. For a general separable Hilbert space $\mathcal{Y}$, let $\mathcal{W} = \mathcal{Y}^m$ and $C_1, \ldots, C_m : \mathcal{Y} \to \mathcal{Y}$ be bounded linear operators. For $f(x) = (f^1(x), \ldots, f^m(x))$, with each $f^i(x) \in \mathcal{Y}$, we can define the operator $C = [C_1, \ldots, C_m] : \mathcal{Y}^m \to \mathcal{Y}$ by

$$Cf(x) = C_1 f^1(x) + \cdots + C_m f^m(x) \in \mathcal{Y}. \quad (23)$$

This gives rise to a vector-valued version of multi-view learning, where outputs from m views, each one being a vector in the Hilbert space $\mathcal{Y}$, are linearly combined.

Consider the following setting of vector-valued multi-view learning, which we apply to the problem of multi-class categorization in Section 5. Let the input space be $\mathcal{X}$, a non-empty subset, and the output space be $\mathcal{Y} = \mathbb{R}^P$, $P \in \mathbb{N}$. Let the number of views be $m \in \mathbb{N}$. For the multi-class classification problem, P is the number of classes. For each y_i, $1 \leq i \leq l$, in the labeled training sample, $y_i = (-1, \ldots, 1, \ldots, -1)$, with 1 at the kth location if x_i is in the kth class.

Let $\mathcal{W} = \mathcal{Y}^m = \mathbb{R}^{Pm}$. The hypothesis space $\mathcal{H}_K$ of functions with values in $\mathcal{W} = \mathbb{R}^{Pm}$ is induced by a positive definite matrix-valued kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{Pm \times Pm}$, that is for each pair $(x, t) \in \mathcal{X} \times \mathcal{X}$, $K(x, t)$ is an $Pm \times Pm$ matrix. For each function $f \in \mathcal{H}_K$, $f(x) = (f^1(x), \ldots, f^m(x))$, where $f^i(x) \in \mathcal{Y} = \mathbb{R}^P$ is the value corresponding to the ith view.

In Proposition 2, $J_l^{\mathcal{W},u+l}$ is a diagonal matrix of size $Pm(u+l) \times Pm(u+l)$, with the first Pml entries on the main diagonal being 1, the rest being 0, $M \in \mathbb{R}^{Pm(u+l) \times Pm(u+l)}$, $K[\mathbf{x}] \in \mathbb{R}^{Pm(u+l) \times Pm(u+l)}$, $C \in \mathbb{R}^{P \times Pm}$, $\mathbf{C} \in \mathbb{R}^{P(u+l) \times Pm(u+l)}$, $\mathbf{C}^*\mathbf{C} \in \mathbb{R}^{Pm(u+l) \times Pm(u+l)}$, $\mathbf{a} \in \mathbb{R}^{Pm(u+l)}$, $\mathbf{y} \in \mathbb{R}^{P(u+l)}$.

4.1. The Combination Operator

In the present context, the bounded linear operator $C : \mathcal{W} \to \mathcal{Y}$ is a matrix of size $P \times Pm$. This operator transforms the output vectors obtained from the m views f^i's in $\mathbb{R}^{Pm}$ into an output vector in $\mathbb{R}^P$. The simplest form of C is the average operator:

$$Cf(x) = \frac{1}{m}(f^1(x) + \cdots + f^m(x)) \in \mathcal{Y} = \mathbb{R}^P. \quad (24)$$

Let $\otimes$ denote the Kronecker tensor product. For $m \in \mathbb{N}$, let $\mathbf{e}_m = (1, \ldots, 1)^T \in \mathbb{R}^m$. The matrix C is then

$$C = \frac{1}{m}\mathbf{e}_m^T \otimes I_P = \frac{1}{m}[I_P, \ldots, I_P]. \quad (25)$$

More generally, we consider a weight vector $\mathbf{c} = (c_1, \ldots, c_m)^T \in \mathbb{R}^m$ and define C as

$$C = \mathbf{c}^T \otimes I_P, \text{ with } Cf(x) = \sum_{i=1}^m c_i f^i(x) \in \mathbb{R}^P. \quad (26)$$

4.2. Multi-view Manifold Regularization

We decompose the multi-view manifold regularization term $\gamma_I \langle \mathbf{f}, M\mathbf{f} \rangle_{\mathcal{W}^{u+l}}$ in (Eq. 9) into two components

$$\gamma_I \langle \mathbf{f}, M\mathbf{f} \rangle_{\mathcal{W}^{u+l}} = \gamma_B \langle \mathbf{f}, M_B \mathbf{f} \rangle_{\mathcal{W}^{u+l}} + \gamma_W \langle \mathbf{f}, M_W \mathbf{f} \rangle_{\mathcal{W}^{u+l}}, \quad (27)$$

where $M_B, M_W : \mathcal{W}^{u+l} \to \mathcal{W}^{u+l}$ are symmetric, positive operators, and $\gamma_B, \gamma_W \geq 0$. We call the first term *between-view regularization*, which measures the consistency of the component functions across different views, and the second term *within-view regularization*, which measures the smoothness of the component functions in their corresponding views. We describe next two choices for M_B and M_W.

Between-view Regularization. Let

$$M_m = mI_m - \mathbf{e}_m \mathbf{e}_m^T. \quad (28)$$

This is the $m \times m$ matrix with $(m-1)$ on the diagonal and -1 elsewhere. Then for $\mathbf{a} = (a_1, \ldots, a_m) \in \mathbb{R}^m$,

$$\mathbf{a}^T M_m \mathbf{a} = \sum_{j,k=1, j<k}^m (a_j - a_k)^2. \quad (29)$$

If each $a_i \in \mathbb{R}^P$, then we have $\mathbf{a} \in \mathbb{R}^{Pm}$ and

$$\mathbf{a}^T (M_m \otimes I_P)\mathbf{a} = \sum_{j,k=1, j<k}^m ||a_j - a_k||^2_{\mathbb{R}^P}. \quad (30)$$

We define M_B by

$$M_B = I_{u+l} \otimes M_{Pm} = I_{u+l} \otimes (M_m \otimes I_P). \quad (31)$$

Then M_B is a diagonal block matrix of size $Pm(u+l) \times Pm(u+l)$, with each block (i,i) being M_{Pm}. For $\mathbf{f} = (f(x_1), \ldots, f(x_{u+l})) \in \mathbb{R}^{Pm(u+l)}$, with $f(x_i) \in \mathbb{R}^{Pm}$,

$$\langle \mathbf{f}, M_B \mathbf{f} \rangle_{\mathbb{R}^{Pm(u+l)}} = \sum_{i=1}^{u+l} \langle f(x_i), M_{Pm} f(x_i) \rangle_{\mathbb{R}^{Pm}}$$

$$= \sum_{i=1}^{u+l} \sum_{j,k=1, j<k}^m ||f^j(x_i) - f^k(x_i)||^2_{\mathbb{R}^P}. \quad (32)$$

For $P = 1$, this is precisely the Point Cloud regularization term for scalar multi-view learning (Rosenberg et al., 2009; Brefeld et al., 2006). In particular, for $m = 2$, we have $M_2 = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$, and

$$\langle \mathbf{f}, M_B\mathbf{f}\rangle_{\mathbb{R}^{2(u+l)}} = \sum_{i=1}^{u+l}(f^1(x_i) - f^2(x_i))^2, \quad (33)$$

which is the Point Cloud regularization term for co-regularization (Sindhwani & Rosenberg, 2008).

Within-view Regularization. One way to define M_W is via the graph Laplacian. For view i, $1 \leq i \leq m$, let G^i be a corresponding undirected graph, with symmetric, nonnegative weight matrix W^i, which induces the scalar graph Laplacian L^i, a matrix of size $(u + l) \times (u + l)$. For a vector $\mathbf{a} \in \mathbb{R}^{u+l}$, we have

$$\mathbf{a}^TL^i\mathbf{a} = \sum_{j,k=1,j<k}^{u+l} W^i_{jk}(a_j - a_k)^2.$$

Let L be the block matrix of size $(u+l)\times(u+l)$, with block (i, j) being the $m \times m$ diagonal matrix given by

$$L_{i,j} = \mathrm{diag}(L^1_{ij}, \dots L^m_{ij}). \quad (34)$$

Then for $\mathbf{a} = (a_1, \dots, a_{u+l})$, with $a_j \in \mathbb{R}^m$, we have

$$\mathbf{a}^TL\mathbf{a} = \sum_{i=1}^{m}\sum_{j,k=1,j<k}^{u+l} W^i_{jk}(a^i_j - a^i_k)^2. \quad (35)$$

If $a_j \in \mathbb{R}^{Pm}$, with $a^i_j \in \mathbb{R}^P$, then

$$\mathbf{a}^T(L \otimes I_P)\mathbf{a} = \sum_{i=1}^{m}\sum_{j,k=1,j<k}^{u+l} W^i_{jk}||a^i_j - a^i_k||^2_{\mathbb{R}^P}. \quad (36)$$

Define

$$M_W = L \otimes I_P, \quad \text{then} \quad (37)$$

$$\langle \mathbf{f}, M_W\mathbf{f}\rangle_{\mathbb{R}^{Pm(u+l)}} = \sum_{i=1}^{m}\sum_{j,k=1,j<k}^{u+l} W^i_{jk}||f^i(x_j) - f^i(x_k)||^2_{\mathbb{R}^P}. \quad (38)$$

The ith summand in the sum $\sum_{i=1}^m$ is precisely a manifold regularization term within view i.

Single View Case. When $m = 1$, $M_m = 0$ and $M_B = 0$. In this case, we simply carry out manifold regularization within the given single view, using M_W.

4.3. Numerical Implementation

This section is devoted to giving an efficient numerical implementation of Proposition 2 while still preserving the closed form of the solution. Assume that each input x is decomposed into $x = (x^1, \dots, x^m)$ for the m different views. We define $K(x, t)$ as a block diagonal matrix, with the (i, i)th block given by

$$K(x, t)_{i,i} = k^i(x^i, t^i)I_P, \quad (39)$$

where k^i is a scalar-valued kernel, such as the Gaussian kernel. Then $K(x, t)$ is a matrix of size $Pm \times Pm$.

Simplification via the Kronecker Tensor Product. Let $G(x, t)$ be the $m \times m$ diagonal matrix, with

$$(G(x, t))_{i,i} = k^i(x^i, t^i), \quad (40)$$

and $G[\mathbf{x}]$ be the $(u + l) \times (u + l)$ block matrix, where block (i, j) is the $m \times m$ matrix $G(x_i, x_j)$. Then

$$K(x, t) = G(x, t) \otimes I_P, \quad (41)$$

and the matrix $K[\mathbf{x}]$ is

$$K[\mathbf{x}] = G[\mathbf{x}] \otimes I_P. \quad (42)$$

Proposition 3. *For $C = \mathbf{c}^T \otimes I_P$, $\mathbf{c} \in \mathbb{R}^m$, $M_W = L \otimes I_P$, $M_B = I_{u+l} \otimes (M_m \otimes I_P)$, the system of linear equations (13) in Proposition 2 is equivalent to*

$$BA = Y_C, \quad (43)$$

where

$$B = \left((J^{u+l}_l \otimes \mathbf{cc}^T) + l\gamma_B(I_{u+l} \otimes M_m) + l\gamma_W L\right) G[\mathbf{x}] + l\gamma_A I_{(u+l)m}, \quad (44)$$

which is of size $(u + l)m \times (u + l)m$, A is the matrix of size $(u + l)m \times P$ such that $\mathbf{a} = \mathrm{vec}(A^T)$, and Y_C is the matrix of size $(u + l)m \times P$ such that $\mathbf{C}^*\mathbf{y} = \mathrm{vec}(Y_C^T)$. $J^{u+l}_l : \mathbb{R}^{u+l} \to \mathbb{R}^{u+l}$ is a diagonal matrix of size $(u+l)\times(u+l)$, with the first l entries on the main diagonal being 1 and the rest being 0.

Evaluation on a Testing Sample. Let $\mathbf{v} = \{v_1, \dots, v_t\} \in \mathcal{X}$ be an arbitrary set of testing input examples, with $t \in \mathbb{N}$. Let $\mathbf{f}_{\mathbf{z},\gamma}(\mathbf{v}) = (\{f_{\mathbf{z},\gamma}(v_1), \dots, f_{\mathbf{z},\gamma}(v_t)\})^T \in \mathbb{R}^{Pmt}$, with

$$f_{\mathbf{z},\gamma}(v_i) = \sum_{j=1}^{u+l} K(v_i, x_j)a_j.$$

Let $K[\mathbf{v}, \mathbf{x}]$ denote the $t \times (u + l)$ block matrix, where block (i, j) is $K(v_i, x_j)$ and similarly, let $G[\mathbf{v}, \mathbf{x}]$ denote the $t \times (u + l)$ block matrix, where block (i, j) is the $m \times m$ matrix $G(v_i, x_j)$. Then

$$\mathbf{f}_{\mathbf{z},\gamma}(\mathbf{v}) = K[\mathbf{v}, \mathbf{x}]\mathbf{a} = (G[\mathbf{v}, \mathbf{x}]\otimes I_P)\mathbf{a} = \mathrm{vec}(A^TG[\mathbf{v}, \mathbf{x}]^T),$$

In particular, for $\mathbf{v} = \mathbf{x} = (x_i)^{u+l}_{i=1}$, the original training sample, we have $G[\mathbf{v}, \mathbf{x}] = G[\mathbf{x}]$.

Algorithm 1 combines all the steps in this section.

Algorithm 1 $\mathbb{R}^P$-valued, m-view, semi-supervised least square regression and classification

```
Input: - Training data $\mathbf{z} = (x_i, y_i)_{i=1}^{l} \cup (x_i)_{i=l+1}^{u+l}$, with $l$ labeled examples and $u$ unlabeled examples.
- Number of views: $m$. Output dimension: $P$.
- Testing example: $v$.
Parameters: $\gamma_A, \gamma_B, \gamma_W$; a kernel $k^i$ for each view and its parameters; the weight vector $\mathbf{c}$.
Procedure: - Compute kernel matrix $G[\mathbf{x}]$ on input set $\mathbf{x} = (x_i)_{i=1}^{u+l}$ according to (40).
- Compute matrix $C$ according to (26).
- Compute graph Laplacian $L$ according to (34).
- Compute matrices $B$, $Y_C$ according to Proposition 3.
- Solve system of linear equations $BA = Y_C$.
- Compute kernel matrix $G[v, \mathbf{x}]$ between $v$ and $\mathbf{x}$.
Output: $f_{\mathbf{z},\gamma}(v) = \mathrm{vec}(A^T G[v, \mathbf{x}]^T) \in \mathbb{R}^{Pm}$.
$\mathbb{R}^P$-regression: return $Cf_{\mathbf{z},\gamma}(v) \in \mathbb{R}^P$.
$P$-way classification: return index of $\max(Cf_{\mathbf{z},\gamma}(v))$.
```

5. Experiments

This section provides an empirical analysis of the proposed framework and its particular instance described in Section 4. We tested our method on two different multi-class categorization tasks, namely object recognition and bird species categorization, using challenging, publicly available datasets (Fei-Fei et al., 2006; Wah et al., 2011). For these problems, each *view* of an input example is one type of features of that example.

These experiments demonstrate that: 1) the multi-view regularization terms both contribute to improve learning performance; 2) multi-view learning achieves significantly better performance compared to single-view learning; 3) our method gives comparable performance with other state-of-the-art methods.

Parameters. These were fixed for all of the experiments unless explicitly stated otherwise. In particular, $\gamma_A = 10^{-5}$, $\gamma_B = \gamma_W = 10^{-6}$, and the weight vector $\mathbf{c}$ of the combination operator was set to be uniform. The graph Laplacians of Eq. 34 were computed with the kernel matrices as weight matrices.

Object Recognition. We evaluated the proposed method on the Caltech-101 dataset (Fei-Fei et al., 2006) using the features, kernel matrices, and evaluation protocol proposed in (Vedaldi et al., 2009), available at `http://www.robots.ox.ac.uk/~vgg/software/MKL/`. The appearance descriptors are made up of 4 features (views): PHOW gray and color, geometric blur (GB), and self-similarity (SSIM). For the first three features, a three-level pyramid of spatial histograms were built (Vedaldi et al., 2009). First, we analyzed the contributions of each of the between-view (Eq. 32) and within-view (Eq. 38) regularization terms in (Eq. 9). A subset of 10 images for each class were randomly selected, with half used as labeled data $l_c = 5$ and the other half as unlabeled data $u_c = 5$ (see Table 1, last column). We also tested the proposed method in the one-shot learning setup, where the number of labeled images is one per class $l_c = 1$ (see Table 1, third column). The testing set consisted of 15 images per category. For this test, we selected the features at the bottom of each pyramid, because they give the best performance in practice. We can see from Table 1 that both the between-view and within-view regularization terms contribute to increase the recognition rate, e.g. with $l_c = 1$ the improvement is 2.35%. As one would expect, the improvement resulting from the use of unlabeled data is bigger when there are more unlabeled data than labeled data, which can be seen by comparing the third and forth columns.

Table 1. Results on the Caltech-101 dataset using PHOW color and gray L2, SSIM L2 and GB. The training set is made of 1 or 5 labeled data l_c and 5 unlabeled data per class u_c, and 15 images per class are left for testing.

		Accuracy	Accuracy
γ_B	γ_W	$l_c = 1, u_c = 5$	$l_c = u_c = 5$
0	0	30.59%	63.68%
0	10^{-6}	31.81%	63.97%
10^{-6}	0	32.44%	64.18%
10^{-6}	10^{-6}	**32.94**%	**64.2**%

We observed that averaging views in the combination operator is not the optimal choice. For $\mathbf{c} = [5, 10, 10, 10]^T$, we obtained 65.22% accuracy, an improvement of 1.02% with respect to the last row, last column of Table 1. This suggests that $\mathbf{c}$ should be either jointly optimized within the proposed framework or cross-validated. We leave this to a future work.

To demonstrate that multi-view learning is able to combine features properly, we report in Table 2 the performance of each feature independently and of the proposed method with all 10 views combined. The minimum improvement with respect to the view that gives the best results (PHOW gray L2) is 4.77% (second column) and 4.71% (last column).

The last test we performed on the Caltech-101 dataset used the features at the bottom level of the pyramid in a supervised setup with 15 images per category for training. The results obtained (see Table 3) are comparable with other state-of-the-art techniques. This is very encouraging, because the loss function V in the present implementation of our framework is the least

Table 2. Results on the Caltech-101 data set using each feature in the single-view learning framework and all 10 features in the multi-view learning framework (last row).

Feature	Accuracy $l_c = 1, u_c = 5$	Accuracy $l_c = u_c = 5$
PHOW color L0	13.66%	33.14%
PHOW color L1	17.1%	42.03%
PHOW color L2	18.71%	45.86%
PHOW gray L0	20.31%	45.38%
PHOW gray L1	24.53%	54.86%
PHOW gray L2	25.64%	56.75%
SSIM L0	15.27%	35.27%
SSIM L1	20.83%	45.12%
SSIM L2	22.64%	48.47%
GB	25.01%	44.49%
10-view learning	**30.41%**	**61.46%**

square loss, while most existing methods use SVM that is known to reach very good classification performance in practice. We leave to future work an implementation using the SVM loss function, which will be able to combine the advantages of our framework with the classification power of the SVM.

Table 3. Comparison with state-of-the-art methods on the Caltech-101 dataset using PHOW color and gray L2, SSIM L2 and GB in the supervised setup . SVM = Support Vector Machine, GP = Gaussian Process, LS = Least Square.

Method	Classifier	Accuracy $l_c = 15$
(Gehler & Nowozin, 2009)	SVM-based	**74.6%**
(Yang et al., 2009)	SVM-based	73.2%
(Christoudias et al., 2009)	GP	73.00%
(Vedaldi et al., 2009)	SVM	71.1%
4-view learning	LS	73.33%

Bird Species Categorization. Next, we examined the Caltech-UCSD Birds-200-2011 dataset (Wah et al., 2011), which is particularly challenging because it contains 11, 788 annotated images of 200 very similar bird species. We consider as two views an appearance-based feature and an attribute-based feature. For the appearance-based view, we extracted the bag of feature descriptors based on PHOW gray features (Vedaldi et al., 2009) and computed the χ^2 kernel. For the attribute-based view, we used the 312-dimensional binary vector provided in (Wah et al., 2011) and computed the Gaussian kernel; these correspond to 312 attributes for each image, which were not exploited as features in (Wah et al., 2011). The parameters of both kernels were set to be the median pairwise distances among training points. Figure 1 shows the classification accuracies of single-view learning (in red and green) of each feature and 2-view learning (in blue) when increasing the number of labeled data in the training set with a fixed number of unlabeled data. The best categorization accuracy reported so far in (Wah et al., 2011) is 6.94% (with $l_c = 5$). Their pipeline, consisting of a bird detector, an appearance-based descriptor and SVM, outperforms PHOW because they performed bird localization while we extracted features from the whole image. The most important fact is that 2-view learning is always able to combine both appearance-based and attribute-based features, giving better results compared to each feature taken as single view.

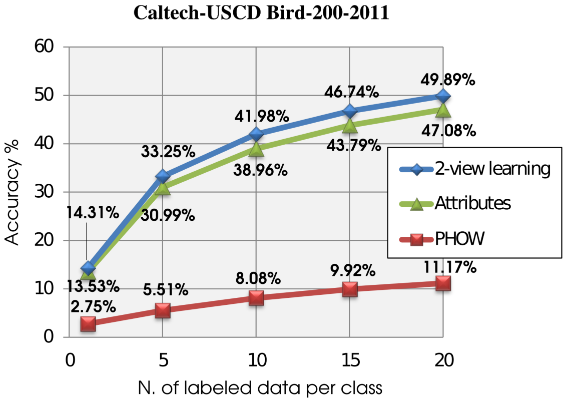

Figure 1. Bird species categorization when increasing the number of labeled data $l_c = \{1, 5, 10, 15, 20\}$ in the training set, with fixed number of unlabeled data $u_c = 5$.

Computational complexity. The system of linear equations (43) in Proposition 3 is of size $m(u + l) \times m(u+l)$, has a unique solution, is simple to implement and is efficient. For the experiments that we report in this paper, the main computation cost is in computing the different kernel matrices. With 4 views, using the same setup of Table 2, the proposed algorithm took 30.79 sec. and 0.48 sec.[2] for training (3060 images) and testing (1530 images), respectively (given the precomputed kernel matrices). With 2 views, it took 4.95 sec. (training) and 0.17 sec. (testing).

6. Conclusion and Future Work

We have presented a general vector-valued RKHS formulation for Semi-Supervised Learning, of which Vector-valued Manifold Regularization and Multi-view Learning are special instances. The results we have obtained demonstrate that this is a promising venue for further research exploration. Our future work includes an implementation of the SVM loss function within our framework, the optimization of the combination operator, and connection to Multiple Kernel Learning.

[2] The method is implemented in MATLAB, on an Intel Xeon(R) CPU E5645 @2.40GHz×12 cores, 12GB RAM.

References

Belkin, M., Niyogi, P., and Sindhwani, V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

Brefeld, U., Gärtner, T., Scheffer, T., and Wrobel, S. Efficient co-regularised least squares regression. In *Proceedings of the International Conference on Machine Learning(ICML)*, 2006.

Brouard, C., D'Alche-Buc, F., and Szafranski, M. Semi-supervised penalized output kernel regression for link prediction. In *Proceedings of the International Conference on Machine Learning*, 2011.

Caponnetto, A., Pontil, M., C.Micchelli, and Ying, Y. Universal multi-task kernels. *Journal of Machine Learning Research*, 9:1615–1646, 2008.

Carmeli, C., Vito, E. De, and Toigo, A. Vector-valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Analysis and Applications*, 4:377–408, 2006.

Christoudias, M., Urtasun, R., and Darrell, T. Bayesian localized multiple kernel learning. *Univ. California Berkeley, Berkeley, CA*, 2009.

Dinuzzo, F., Ong, C.S., Gehler, P., and Pillonetto, G. Learning output kernels with block coordinate descent. In *ICML*, 2011.

Fei-Fei, L., Fergus, R., and Perona, P. One-shot learning of object categories. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(4):594 –611, april 2006.

Gehler, P. and Nowozin, S. On feature combination for multiclass object classification. In *Proceedings of the 12th IEEE International Conference on Computer Vision*, 2009.

Kadri, H., Rabaoui, A., Preux, P., Duflos, E., and Rakotomamonjy, A. Functional regularized least squares classification with operator-valued kernels. In *Proceedings of the International Conference on Machine Learning(ICML)*, 2011.

Micchelli, C. A. and Pontil, M. On learning vector-valued functions. *Neural Computation*, 17:177–204, 2005.

Minh, H.Q. and Sindhwani, V. Vector-valued manifold regularization. In *ICML*, 2011.

Reisert, M. and Burkhardt, H. Learning equivariant functions with matrix valued kernels. *J. Mach. Learn. Res.*, 8:385–408, 2007.

Rosenberg, D., Sindhwani, V., Bartlett, P., and Niyogi, P. A kernel for semi-supervised learning with multi-view point cloud regularization. *IEEE Sig. Proc. Mag.*, 26(5):145–150, 2009.

Saffari, A., Leistner, C., Godec, M., and Bischof, H. Robust multi-view boosting with priors. In *Proceedings of the European Conference on Computer Vision*, 2010.

Schölkopf, B. and Smola, A. *Learning with kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, Cambridge, 2002.

Shawe-Taylor, J. and Cristianini, N. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

Sindhwani, V. and Rosenberg, D. An RKHS for multi-view learning and manifold co-regularization. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2008.

Vedaldi, A., Gulshan, V., Varma, M., and Zisserman, A. Multiple kernels for object detection. In *Proceedings of 12th IEEE International Conference on Computer Vision*, 2009.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The Caltech-UCSD Birds-200-2011 dataset. 2011. URL http://www.vision.caltech.edu/visipedia/CUB-200-2011.html.

Yang, J., Li, Y., Tian, Y., Duan, L., and Gao, W. Group-sensitive multiple kernel learning for object categorization. In *Proceedings of the 12th IEEE International Conference on Computer Vision*, 2009.