Regularized Semipaired Kernel CCA for Domain Adaptation

Siamak Mehrkanoon and Johan A.K. Suykens

***Abstract*—Domain adaptation learning is one of the fundamental research topics in pattern recognition and machine learning. This paper introduces a Regularized Semi-Paired Kernel Canonical Correlation Analysis (RSP-KCCA) formulation for learning a latent space for the domain adaptation problem. The optimization problem is formulated in the primal-dual LS-SVM setting where side information can be readily incorporated through regularization terms. The proposed model learns a joint representation of the data set across different domains by solving a generalized eigenvalue problem or linear system of equations in the dual. The approach is naturally equipped with out-of-sample extension property which plays an important role for model selection. Furthermore, the Nyström approximation technique is used to make the computational issues due to the large size of the matrices involved in the eigendecomposition feasible. The learnt latent space of the source domain is fed to a Multi-Class Semi-Supervised Kernel Spectral Clustering model, MSS-KSC, that can learn from both labeled and unlabeled data points of the source domain in order to classify the data instances of the target domain. Experimental results are given to illustrate the effectiveness of the proposed approaches on synthetic and real-life datasets.**

***Index Terms*—Semi-supervised learning, domain adaption, Kernel Canonical Correlation Analysis, Nyström Approximation**

I. Introduction

THE most common underlying assumption of many machine learning algorithms is that both training and test data exhibit the same distribution or same feature domains. However in many cases the data change from one domain to another or its statistical properties evolve in time. The non-stationary nature of the data brings a new challenge for many existing learning algorithms, which are based on the stationary assumption. When there is a distributional, feature space and/or dimension mismatch between the two domains, the models learned with data in one domain would fail to predict the test data in the other.

In practice, collecting training data in different domains is costly, therefore domain adaptation algorithms seek to generalize a model trained in a source domain (training domain) to a new target domain (test domain). In many practical cases, the source and target distributions can differ substantially. For example, consider the cross-domain image-based object recognition where the acquired images show systematic difference in two setups due to the change of light, sensors and other conditions which implies that the source and target domains are from different image modalities. Taking cross-language document classificatio as an example, documents in English do not share the same feature representation with those in German due to different vocabularies. Another example is for instance, in sentiment analysis problem, where the task is to classify if a Book-review (source domain) or DVD-review (target domain) is positive or negative.

Depending on the availability of the labeled instances in both domains, three scenarios can be considered, i.e. unsupervised, supervised and semi-supervised domain adaptation models. Unsupervised domain adaptation approaches, do not take label information into consideration when learning the feature representation [1]–[3]. On the other hand, supervised domain adaptation approaches, only use labeled data from the source and target domains [4]. In the semi-supervised setting, one learns from labeled source instances as well as a small fraction of the target labeled instances [5], [6]. This setting can have many real world applications, as collecting labeled instances might be costly.

There are three major directions of work being proposed in the literature to address the domain adaptation problems: feature transformation [1], [2], [7]–[9], sample reweighting [10], [11] and dynamic ensemble [12], [13]. In the sample reweighting approach, one assigns sample-dependent weights to the training data with the aim of minimizing the distribution discrepancy between the source and target data points in the reweighted space [14], [15]. The mechanisms that are mostly used in the literature for the estimation of sample dependent weights are formulated as the density ratio between the probability densities of the two domains [14].

Another key research challenge in domain adaptation is how to learn a domain-invariant feature representation for both source and target domains. The adaptation then can be accomplished by learning a model on the new space. The authors in [7] introduced a method to learn the feature transformation in order to produce a set of common transfer components across domains. The Structural Correspondence Learning method proposed in [2] learns a common feature space by identifying correspondence among features from different domains. A domain adaptation approach that uses the correlation subspace as a joint representation of the source and target data is introduced in [9]. In this approach the new representation is learnt using unlabeled data pairs in both source and target domains. A deep learning approach to learn new cross-domain feature representation from the source and target data is proposed in [8].

In many domain adaptation problems, side information in the form of correspondence instances (paired instances) are available for either unlabeled or labeled instances across do-

S. Mehrkanoon and J.A.K. Suykens are with the Department of Electrical Engineering ESAT-STADIUS, KU Leuven, B-3001 Leuven, Belgium. (e-mail:mehrkanoon2011@gmail.com, {siamak.mehrkanoon,johan.suykens}@esat.kuleuven.be).

Manuscript received January, 2016.

mains. For instance consider the snapshots of the actions of the same persons in two different time instance shown in Fig. 1. In this case one can have access to paired labeled and unlabeled samples across domains. In general the instance similarity constraints between domains, if available, can be used to enhance the performance of the classif er [6]. The authors in [16] proposed a method that adapts representations using a small number of paired synthetic and real views of the same object/scene. In their experiments, each real example is paired with a corresponding synthetic image in the same pose, and an additional unpaired synthetic images are also provided as training data.

In general, two types of domain adaptation problems have been addressed in the literature, i.e homogeneous and heterogeneous domain adaption. However, domain adaptation across heterogeneous feature spaces, in which the distributions, feature domains or feature dimensions in source and target domains are different, has gained more attention recently due to its importance in many real life problems [17], [18].

It is the purpose of this paper to present a two step approach for the domain adaption problem. The proposed approach belongs to the feature transformation methods. In the f rst step we formulate a primal-dual optimization problem with the LSSVM-based kernel canonical correlation analysis (LSSVM-KCCA) [19] as core model for learning a new latent space for the domain adaptation problem. Thanks to the primal-dual formulation, as in classical support vector machines, side information is incorporated in the primal model via adding regularization terms and the optimal model representation is obtained in dual. In this way, different scenarios relevant to the domain adaptation problems can be considered and accordingly suitable regularization terms can be adopted. In particular, here we focus on three cases, (see Fig. 2), where we have access to (a) unlabeled paired cross-domain data, (b) unlabeled semi-paired cross-domain data, (c) unlabeled as well as partially labeled paired cross-domain data. Here instances that have correspondences in target domain are referred to as paired otherwise unpaired instances. An example of paired instances is when the same object is tracked between video frames, see Fig. 1. In addition, we assume that the paired instances share the same labels in the source and target domains. Three novel formulations, with LSSVM-KCCA core model, corresponding to the above-mentioned cases are developed to learn a new representation of the source and target data. In addition, in order to make the proposed formulations scalable, the Nyström approximation technique is used to provide an explicit feature map and the corresponding optimization problems are solved in the primal. For the f rst two cases (a) and (b), one needs to solve a generalized eigenvalue problem in dual. Whereas in the third case (c), by incorporating the labels information to the optimization problem, a linear system of equation is solved in the dual. After the new representation are learned, second step consists of learning a completely supervised model (if enough labeled instances are available) or a semi-supervised model that can learn from both labeled and unlabeled data points using the projected training data points. The models are then tested on the unseen test data points.

This paper is organized as follows. In Section II a brief review about Least squares support vector machines based kernel canonical component analysis is given. Section III, brief y reviews the Nyström method for approximating the f nite dimensional feature map. In section IV, we formulate our *Fixed-Size KCCA* for learning a joint representation where unlabeled paired cross-domain data are available. The *Semi-Paired KCCA* (SP-KCCA) formulation is introduced in section V for the case where one only has access to semi-paired unlabeled cross-domain data. In Section VI, the *Regularized Semi-Paired KCCA* (RSP-KCCA) model is proposed where the available labels are integrated into the primal optimization problem via a regularization term and the model is learned using both labeled and unlabeled paired instances. The model selection as well as the experiments results are discusses in Section VII and VIII respectively. The numerical experiments, on both synthetic and real-life datasets with heterogeneous features in different application domains, are carried out to demonstrate the viability of the proposed methods.



Fig. 1. Example of labeled and unlabeled paired instances. Objects that are present in both domains are paired instances. Among paired instances, some of them can be labeled as well. In this illustrative example of two domains, the paired labels are jumping and check-watch actions.

II. OVERVIEW OF LS-SVM BASED KERNEL CCA

Canonical correlation analysis (CCA) is a dimensionality reduction technique for paired data. The problem of CCA consists of measuring the linear relationship between two pairs of variables. Let $\mathcal{D}^{(1)}$ and $\mathcal{D}^{(2)}$ with

$$\mathcal{D}^{(1)} = \{ \underbrace{x_1^{(1)}, ..., x_{n_p}^{(1)}}_{\substack{paired\ and\ unlabeled \\ (\mathcal{D}_{p,ul}^{(1)})}} \},$$

and

$$\mathcal{D}^{(2)} = \{ \underbrace{x_1^{(2)}, ..., x_{n_p}^{(2)}}_{\substack{paired\ and\ unlabeled \\ (\mathcal{D}_{p,ul}^{(2)})}} \},$$

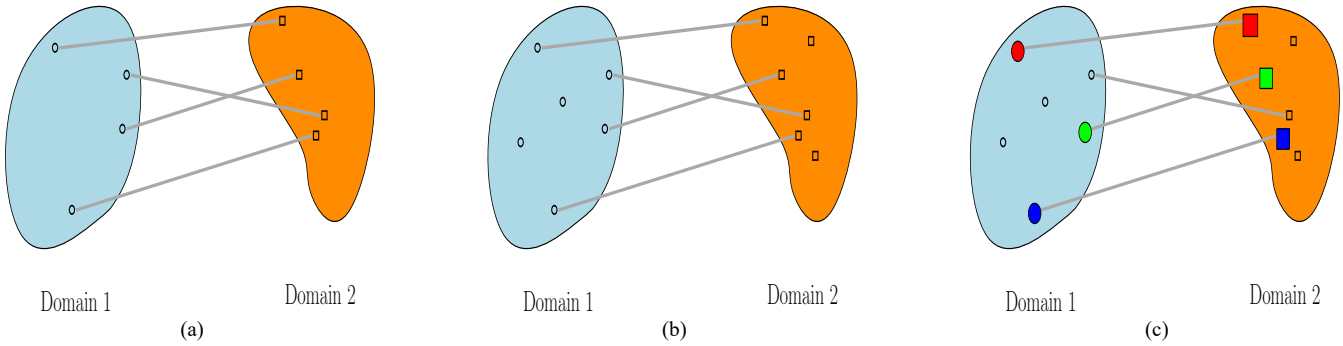


Fig. 2. Illustration of three scenarios studied in this paper. **(a)** Unlabeled paired cross domain data points. (The formulation is explained in Section 4). **(b)** Unlabeled and semi-paired cross domain data points. (The formulation is explained in Section 5). **(c)** Unlabeled + partially labeled paired cross-domain data. (The formulation is explained in Section 6).

denote n_p training observations of $x^{(1)}$ and $x^{(2)}$ from two domains. The objective of CCA is to find basis vectors $w^{(1)}$ and $w^{(2)}$ such that the projected variables ${w^{(1)}}^T x^{(1)}$ and ${w^{(2)}}^T x^{(2)}$ are maximally correlated:

$$\max_{w^{(1)},w^{(2)}} \rho = \frac{{w^{(1)}}^T C_{12} w^{(1)}}{\sqrt{{w^{(1)}}^T C_{11} w^{(1)}}\sqrt{{w^{(2)}}^T C_{22} w^{(2)}}} \tag{1}$$

where $C_{11} = \frac{1}{N}\sum_{i=1}^{n_p} x_i^{(1)} {x_i^{(1)}}^T$, $C_{12} = \frac{1}{n_p}\sum_{i=1}^{n_p} x_i^{(1)} {x_i^{(2)}}^T$ and $C_{22} = \frac{1}{n_p}\sum_{i=1}^{n_p} x_i^{(2)} {x_i^{(2)}}^T$ are the covariance matrices. The Equation (1) is typically reformulated as a constrained optimization problem whose solution can be obtained by solving the following generalized eigenvalue problem:

$$\left[\begin{array}{c|c} 0 & C_{12} \\ \hline C_{21} & 0 \end{array}\right] \left[\begin{array}{c} w^{(1)} \\ \hline w^{(2)} \end{array}\right] = \rho \left[\begin{array}{c|c} C_{11} & 0 \\ \hline 0 & C_{22} \end{array}\right] \left[\begin{array}{c} w^{(1)} \\ \hline w^{(2)} \end{array}\right]. \tag{2}$$

Canonical correlation analysis (CCA) finds linear relationships between two sets of variables. The objective is to find basis vectors on which the projected variables are maximally correlated.

A nonlinear extension of CCA using kernels was introduced in [20], [21]. Kernel CCA, on the other hand, first maps the data into a high dimensional feature space induced by a kernel and then performs linear CCA where nonlinear relationships can be found [21]. It should be noticed that without regularization, kernel CCA does not characterize the canonical correlation of the variables and suffers from ill-conditioning. In the literature, several *ad-hoc* regularization schemes havef been proposed to alleviate the above-mentioned problem.

Least squares support vector machines (LS-SVM) based formulations to different problems such as kernel principal component analysis and kernel canonical correlation analysis have been discussed in [19]. Thanks to the primal-dual approach (typical of LS-SVM), the solution in the primal is expressed in terms of the feature map and the regularization terms are incorporated in the primal optimization problem in a natural way. The optimal representation of the model in the dual is obtained which satisfies the KKT optimality conditions. An LS-SVM formulation to kernel CCA was introduced in [19] with the following primal form:

$$\begin{aligned} \max_{w^{(1)},w^{(2)},e,r} \quad & \mu e^T r - \frac{\gamma_1}{2} e^T e - \frac{\gamma_2}{2} r^T r - \frac{1}{2}{w^{(1)}}^T w^{(1)} \\ & - \frac{1}{2}{w^{(2)}}^T w^{(2)} \\ \text{subject to} \quad & e = \Phi_c^{(1)} w^{(1)}, \\ & r = \Phi_c^{(2)} w^{(2)}, \end{aligned} \tag{3}$$

where

$$\Phi_c^{(1)} = [\varphi^{(1)}(x_1^{(1)}) - \hat{\mu}_{\varphi^{(1)}}, \ldots, \varphi^{(1)}(x_{n_p}^{(1)}) - \hat{\mu}_{\varphi^{(1)}}]$$

and

$$\Phi_c^{(2)} = [\varphi^{(2)}(x_1^{(2)}) - \hat{\mu}_{\varphi^{(2)}}, \ldots, \varphi^{(2)}(x_{n_p}^{(2)}) - \hat{\mu}_{\varphi^{(2)}}]$$

are centered feature map matrices with $\hat{\mu}_{\varphi^{(1)}} = \frac{1}{n_p}\sum_{i=1}^{n_p} \varphi^{(1)}(x_i^{(1)})$ and $\hat{\mu}_{\varphi^{(2)}} = \frac{1}{n_p}\sum_{i=1}^{n_p} \varphi^{(2)}(x_i^{(2)})$. Here $\varphi^{(1)}(\cdot) : \mathbb{R}^{d_1} \to \mathbb{R}^{h_1}$ and $\varphi^{(2)}(\cdot) : \mathbb{R}^{d_2} \to \mathbb{R}^{h_2}$ where h_1 and h_2 are the dimensions of the feature space which can be infinite dimensional.

Proposition II.1. *[19] Given two positive definite kernel functions $K_1(s,t) = \varphi^{(1)}(s)^T\varphi^{(1)}(t)$, $K_2(s,t) = \varphi^{(2)}(s)^T\varphi^{(2)}(t)$ and regularization constants γ_1, γ_2 and μ, the solution to the optimization problem (3) is obtained by solving the following generalized eigenvalue problem:*

$$\begin{aligned} &\left[\begin{array}{c|c} 0 & \Omega_c^{(2)} \\ \hline \Omega_c^{(1)} & 0 \end{array}\right] \left[\begin{array}{c} \alpha \\ \hline \beta \end{array}\right] = \\ &\lambda \left[\begin{array}{c|c} \gamma_1\Omega_c^{(1)} + I_N & 0 \\ \hline 0 & \gamma_2\Omega_c^{(2)} + I_N \end{array}\right] \left[\begin{array}{c} \alpha \\ \hline \beta \end{array}\right] \end{aligned} \tag{4}$$

where $\lambda = \frac{1}{\mu}$ and α, β are Lagrange multiplier vectors. $\Omega_c^{(1)}$ and $\Omega_c^{(2)}$ are the centered kernel matrix calculated as $M_c\Omega^{(1)}M_c$ and $M_c\Omega^{(2)}M_c$ with centering matrix $M_c = I_{n_p} - \frac{1}{n_p}1_{n_p}1_{n_p}^T$. Here $\Omega^{(1)}$ and $\Omega^{(2)}$ are kernel matrix whose i,j-th elements are calculated as $\Omega_{i,j}^{(1)} = \varphi^{(1)}(x_i^{(1)})^T\varphi^{(1)}(x_j^{(1)})$ and $\Omega_{i,j}^{(2)} = \varphi^{(2)}(x_i^{(2)})^T\varphi^{(2)}(x_j^{(2)})$.

Note that the regularization is incorporated in the primal optimization problem (3) in a natural way, moreover equation (3)

does not have the same expression as the correlation coeff cient but the numerator and denominator are taken with a positive and negative sign respectively. The projections of the mapped training data onto the i-th eigenvector become:

$$\begin{cases} z_e^{(i)} = \Phi_c^{(1)} w^{(1)} = \Omega_c^{(1)} \alpha^{(i)}, \ i = 1, \ldots, 2n_p \\ z_r^{(i)} = \Phi_c^{(2)} w^{(2)} = \Omega_c^{(2)} \beta^{(i)}, \ i = 1, \ldots, 2n_p. \end{cases} \tag{5}$$

III. APPROXIMATION TO THE FEATURE MAP

For large-scale problems, the cost of storing as well as computing the eigendecompostion of (4) can be prohibitive due to the size of the matrices involved. Alternatively one can f nd a low rank approximation of the kernel matrices for instance by means of Incomplete Cholesky Factorization [22], Nyström method [23], randomized low-dimensional feature space [24], [25] or reduced kernel technique [26], [27]. Nyström approximation technique has been previously successfully applied in kernel methods for large scale classif cation, regression and semi-supervised learning problems [19], [27]. Following our previous research works, here we also propose to use the Nyström approximation method which provides a f nite dimensional feature map that can then be used to solve the optimization problem in the primal. However it should be mentioned that in principle any of the above-mentioned techniques can as well be employed.

Assume that we have a training dataset $\{x_1, \ldots, x_N\}$. Explicit expression for φ can be obtained by means of an eigenvalue decomposition of the kernel matrix Ω. When the size of the training dataset is large, the so called f xed-size approach [19], where the feature map is approximated by the Nyström method [23], [28], can be used. In what follows, we brief y summarize the f xed-size approach.

Consider the Fredholm integral equation of the f rst kind:

$$\int_C K(x, x_j) \varphi_i(x) p(x) dx = \lambda_i \varphi_i(x_j) \tag{6}$$

where C is a compact subset of $\mathbb{R}^d$. The approximation of the eigenfunction $\varphi_i(x)$ in (6) can be obtained by the Nyström method which applies a quadrature rule for discretizing the left-hand side of (6). This will lead to the eigenvalue problem [23]:

$$\frac{1}{n} \sum_{k=1}^{N} K(x_k, x_j) u_{ik} = \lambda_i^{(s)} u_{ij} \tag{7}$$

where the eigenvalues λ_i and eigenfunctions φ_i from the continuous problem (6) can be approximated by the sample eigenvalues $\lambda_i^{(s)}$ and eigenvectors u_i. Therefore, the i-th component of the n-dimensional feature map $\hat{\varphi} : \mathbb{R}^d \to \mathbb{R}^N$, for any point $x \in \mathbb{R}^d$, can be obtained as follows:

$$\hat{\varphi}_i(x) = \frac{1}{\sqrt{\lambda_i^{(s)}}} \sum_{k=1}^{N} u_{ki} K(x_k, x) \tag{8}$$

where $\lambda_i^{(s)}$ and u_i are eigenvalues and eigenvectors of the kernel matrix $\Omega_{N \times N}$. Furthermore, the k-th element of the i-th eigenvector is denoted by u_{ki}. In practice when N is large, we work with a subsample (prototype vectors) of size $m \ll N$ whose elements are selected using an entropy based criterion. In this case, the m-dimensional feature map $\hat{\varphi} : \mathbb{R}^d \to \mathbb{R}^m$ can be approximated as follows:

$$\hat{\varphi}(x) = [\hat{\varphi}_1(x), \ldots, \hat{\varphi}_m(x)]^T \tag{9}$$

where

$$\hat{\varphi}_i(x) = \frac{1}{\sqrt{\lambda_i^{(s)}}} \sum_{k=1}^{m} u_{ki} K(x_k, x), i = 1, \ldots, m \tag{10}$$

where $\lambda_i^{(s)}$ and u_i are now eigenvalues and eigenvectors of the constructed kernel matrix $\Omega_{m \times m}$ using the selected prototype vectors.

We aim at using an m-dimensional approximation to the feature map φ. Therefore we need to select a subset of f xed size m from a pool of training points of size N. As it has been motivated in [19], the Rényi entropy criterion [29] is used, to select m points from training dataset. Once the subset is available, the m-dimensional feature map is obtained using equation (10). It should be noted that m is a user def ned parameter that can be designed in accordance with the available memory of the computer that is being used to conduct the experiments.

IV. FIXED-SIZE KCCA

Given the m-dimensional approximation to the feature map matrices, i.e.

$$\begin{aligned} \hat{\Phi}^{(1)} &= [\hat{\varphi}^{(1)}(x_1^{(1)}), \ldots, \hat{\varphi}^{(1)}(x_{n_p}^{(1)})]^T \in \mathbb{R}^{n_p \times m_1} \\ \hat{\Phi}^{(2)} &= [\hat{\varphi}^{(2)}(x_1^{(2)}), \ldots, \hat{\varphi}^{(2)}(x_{n_p}^{(2)})]^T \in \mathbb{R}^{n_p \times m_2}, \end{aligned} \tag{11}$$

we formulate the Fixed-Size Kernel canonical correlation analysis (FS-KCCA) in primal as follows:

$$\begin{aligned} \max_{w^{(1)}, w^{(2)}, \hat{e}, \hat{r}} \quad & \mu e^T r - \frac{1}{2} {w^{(1)}}^T w^{(1)} - \frac{1}{2} {w^{(2)}}^T w^{(2)} \\ & - \frac{\gamma_1}{2} \hat{e}^T \hat{e} - \frac{\gamma_2}{2} \hat{r}^T \hat{r} \\ \text{subject to} \quad & \hat{e} = \hat{\Phi}_c^{(1)} w^{(1)} \\ & \hat{r} = \hat{\Phi}_c^{(2)} w^{(2)}. \end{aligned} \tag{12}$$

Here $\hat{\Phi}_c^{(1)}$ and $\hat{\Phi}_c^{(2)}$ are centered feature map matrices obtained by subtracting the means of the training samples in the feature space.

Lemma IV.1. *Given a centered finite dimensional approximation to the feature map matrices, $\hat{\Phi}_c^{(1)}$ and $\hat{\Phi}_c^{(2)}$, and regularization constants $\gamma_1, \gamma_2 \in \mathbb{R}^+$, the solution to (12) is obtained by solving the following generalized eigenvalue problem:*

$$\left[\begin{array}{c|c} 0 & (\hat{\Phi}_c^{(1)})^T \hat{\Phi}_c^{(2)} \\ \hline (\hat{\Phi}_c^{(2)})^T \hat{\Phi}_c^{(1)} & 0 \end{array}\right] \left[\begin{array}{c} w^{(1)} \\ \hline w^{(2)} \end{array}\right] =$$

$$\lambda \left[\begin{array}{c|c} \gamma_1 (\hat{\Phi}_c^{(1)})^T (\hat{\Phi}_c^{(1)}) + I_{m_1} & 0 \\ \hline 0 & \gamma_2 (\hat{\Phi}_c^{(2)})^T (\hat{\Phi}_c^{(2)}) + I_{m_2} \end{array}\right] \left[\begin{array}{c} w^{(1)} \\ \hline w^{(2)} \end{array}\right], \tag{13}$$

if μ is selected as $\mu = \frac{1}{\lambda}$.

Proof. Eliminating the constrained from (12) and taking the derivative of the resulting cost function with respect to $w^{(1)}$ and $w^{(2)}$ yields:

$$\begin{cases} \frac{\partial \mathcal{J}}{\partial w^{(1)}} = 0 \Rightarrow \\ (\hat{\Phi}_c^{(1)})^T \hat{\Phi}_c^{(2)} w^{(2)} = \lambda \Big[\gamma_1 (\hat{\Phi}_c^{(1)})^T \hat{\Phi}_c^{(1)} + I_{m_1} \Big] w^{(1)}, \\ \frac{\partial \mathcal{L}}{\partial w^{(2)}} = 0 \Rightarrow \\ (\hat{\Phi}_c^{(2)})^T \hat{\Phi}_c^{(1)} w^{(1)} = \lambda \Big[\gamma_2 (\hat{\Phi}_c^{(2)})^T \hat{\Phi}_c^{(2)} + I_{m_2} \Big] w^{(2)}, \end{cases} \tag{14}$$

which then can be rewritten as in (13). Here $\lambda = \frac{1}{\mu}$. □

Note that in (13), the matrix involved in the eigen-decomposition is of size $(m_1 + m_2) \times (m_1 + m_2)$, where $m_1, m_2 \ll n_p$. The score variables for training data can be expressed as follows:

$$\begin{cases} \hat{z}_e^{(i)} = \hat{\Phi}_c^{(1)} w_i^{(1)} \; i = 1, \ldots, m_1 + m_2 \\ \hat{z}_r^{(i)} = \hat{\Phi}_c^{(2)} w_i^{(2)} \; i = 1, \ldots, m_1 + m_2. \end{cases} \tag{15}$$

Similarly, one can compute the score variables of the new unseen test data points by projecting the centered explicit feature map of the test points onto the learned eigenvectors $w^{(1)}$ and $w^{(2)}$.

Remark IV.1. *In the formulation (3), where an implicit feature map is used, one solves a generalized eigenvalue problem of size $4 \times n_p \times n_p$ in the dual. Therefore the algorithm has $\mathcal{O}((2n_p)^3)$ training complexity with naive implementations. On the other hand, when an explicit feature map is used one can solve the problem in the primal. Given $2n_p$ training points, the complexity of calculating the Nyström approximation, with eigenvalue decomposition of the kernel matrix of size m and solving a generalized eigenvalue problem of size $m \times m$ are is $\mathcal{O}(m^3 + n_p m^2)$ and $\mathcal{O}(m^3)$ respectively where here $m = m_1 + m_2$ and $m \ll n_p$. The total complexity of the proposed method, neglecting lower order terms, is then given by the sum of the two complexities.*

V. Semi-Paired KCCA

The Fixed-size KCCA (FS-KCCA) approach introduced in the previous section IV learns representations that are more closely related to the underlying patterns of the data rather than noise by finding directions that maximize correlation. Although the learnt directions have high correlation, they do not necessarily capture the manifold structure of the data. Furthermore, FS-KCCA requires the data to be paired or one-to-one correspondence among different modalities. However, in real-life applications, pairing data in different modalities can be costly and time consuming whereas the availability of unpaired data are relatively easier. Therefore in practice, normally one encounters semi-paired data i.e. large amount of unpaired and few paired data points.

In the literature, attempts have been made to use the side information hidden in additional unpaired data when finding the canonical directions. SemiCCA proposed in [30] incorporate additional unpaired samples by smoothly bridges the eigenvalue problems of CCA and principal component analysis (PCA). The authors in [31] proposed semi-supervised Laplacian regularization of kernel canonical correlation (SemiLRKCC) to find a set of highly correlated directions by exploiting the intrinsic manifold geometry structure of all data including paired and unpaired data. The method then can use the side information hidden in additional unpaired data. It should be noted that in their terminology, the term "semi-supervised" is referred to the case where "semi-paired" data points is used in their analysis. Several extensions and other variants of this approach have been introduced in the literature see [32]. However, as in SemiLRKCC, these schemes start from an unregularized kernel CCA problem and perform regularization afterwards in an *ad-hoc* manner with no concrete optimization objective.

A. Implicit Feature Map

As already discussed, thanks to the primal-dual formulation, we start by formulating an optimization problem and the regularization terms can naturally be incorporated in the primal optimization problem providing a better insights to the problem [33]–[35]. An optimal representation of the model is then obtained in the dual where the solution meets the KKT optimality conditions. In what follows we formulate a regularized KCCA where the deviation of the projections with respect to the data manifold is penalized. The additional term enforces the smoothness of the solutions with respect to the empirical data manifold. We assume that we are given two training datasets $\mathcal{D}^{(1)}$ and $\mathcal{D}^{(2)}$ from source and target domains with

$$\mathcal{D}^{(1)} = \{ \underbrace{x_1^{(1)}, ..., x_{n_p}^{(1)}}_{\substack{paired \\ (unlabeled) \\ (\mathcal{D}_{p,ul}^{(1)})}}, \underbrace{x_{n_p+1}^{(1)}, .., x_{n_1}^{(1)}}_{\substack{unpaired \\ (unlabeled) \\ (\mathcal{D}_{up,ul}^{(1)})}} \},$$

and

$$\mathcal{D}^{(2)} = \{ \underbrace{x_1^{(2)}, ..., x_{n_p}^{(2)}}_{\substack{paired \\ (unlabeled) \\ (\mathcal{D}_{p,ul}^{(2)})}}, \underbrace{x_{n_p+1}^{(2)}, .., x_{n_2}^{(2)}}_{\substack{unpaired \\ (unlabeled) \\ (\mathcal{D}_{up,ul}^{(2)})}} \},$$

where $\mathcal{D}_{p,ul}^{(1)}$ and $\mathcal{D}_{p,ul}^{(2)}$ are paired samples whereas $\mathcal{D}_{up,ul}^{(1)}$ and $\mathcal{D}_{up,ul}^{(2)}$ are unpaired samples. In other words, the first n_p data points have correspondences in other modality, whereas the remaining instances do not have correspondences. In order to learn the canonical directions, also taking into account the information of unpaired data instances, we formulate the Semi-Paired kernel canonical correlation analysis (SP-KCCA) with implicit feature map matrices $\Phi_c^{(1)}$ and $\Phi_c^{(2)}$ in the primal as follows:

$$\begin{aligned}\max_{w^{(1)},w^{(2)},e,r} \quad & \mu e^T A r - \frac{1}{2}{w^{(1)}}^T w^{(1)} - \frac{1}{2}{w^{(2)}}^T w^{(2)} \\ & - \frac{1}{2}e^T\Big(\gamma_1 P_1 + (1-\gamma_1)L_1\Big)e \\ & - \frac{1}{2}r^T\Big(\gamma_2 P_2 + (1-\gamma_2)L_2\Big)r \\ \text{subject to} \quad & e = \Phi_c^{(1)} w^{(1)}, \\ & r = \Phi_c^{(2)} w^{(2)}.\end{aligned} \tag{16}$$

Here L_1 and L_2 are the graph Laplacian matrices associated with each of the dataset $\mathcal{D}^{(1)}$ and $\mathcal{D}^{(2)}$. Given n_1 training data points from set $\mathcal{D}^{(1)}$ and their corresponding similarity matrix S, the symmetric normalized Laplacian matrix L_1 can be computed as $D^{-1/2}(D-S)D^{-1/2}$ where D is def ned as the diagonal matrix with the degrees $d_i = \sum_{j=1}^{n_1} S_{i,j}$, $i = 1, \ldots, n_1$ on the main diagonal [36]. Similarly, one can compute the Laplacian matrix L_2 for the data points in $\mathcal{D}^{(2)}$. Here matrix P_1 and P_2 are def ned as follows:

$$P_1 = \left[\begin{array}{c|c} I_{n_p \times n_p} & 0_{n_p \times n_1 - n_p} \\ \hline 0_{n_1 - n_p \times n_p} & 0_{n_1 - n_p \times n_1 - n_p} \end{array}\right]_{n_1 \times n_1},$$

$$P_2 = \left[\begin{array}{c|c} I_{n_p \times n_p} & 0_{n_p \times n_2 - n_p} \\ \hline 0_{n_2 - n_p \times n_p} & 0_{n_2 - n_p \times n_2 - n_p} \end{array}\right]_{n_2 \times n_2}.$$

Matrix A is also def ned as follows:

$$A = \left[\begin{array}{c|c} I_{n_p \times n_p} & 0_{n_p \times n_2 - n_p} \\ \hline 0_{n_1 - n_p \times n_p} & 0_{n_1 - n_p \times n_2 - n_p} \end{array}\right]_{n_1 \times n_2},$$

where $I_{n_p \times n_p}$ is the identity matrix of size $n_p \times n_p$ and $0_{s \times t}$ is zero matrix of size $s \times t$.

The fourth and f fth terms in the cost function of (16), encourage smoothness over each one of the score variables e and r so that they do not change too much between nearby points. This kind of regularization terms has been previously employed in other contexts such as spectral clustering [36], [37].

Lemma V.1. *Given two positive definite kernel functions $K_1(s,t) = \varphi^{(1)}(s)^T\varphi^{(1)}(t)$, $K_2(s,t) = \varphi^{(2)}(s)^T\varphi^{(2)}(t)$ and regularization constants γ_1, γ_2 and μ, the solution to the optimization problem (3) is obtained by solving the following generalized eigenvalue problem:*

$$\begin{aligned}&\left[\begin{array}{c|c} 0 & A\Omega_c^{(2)} \\ \hline A^T\Omega_c^{(1)} & 0 \end{array}\right]\left[\begin{array}{c} \alpha \\ \hline \beta \end{array}\right] = \\ &\lambda\left[\begin{array}{c|c} V_1\Omega_c^{(1)} + I_{n_1} & 0 \\ \hline 0 & V_2\Omega_c^{(2)} + I_{n_2} \end{array}\right]\left[\begin{array}{c} \alpha \\ \hline \beta \end{array}\right],\end{aligned} \tag{17}$$

where $V_1 = \gamma_1 P_1 + (1-\gamma_1)L_1$, $V_2 = \gamma_2 P_2 + (1-\gamma_2)L_2$, $\lambda = \frac{1}{\mu}$ and α, β are Lagrange multiplier vectors. $\Omega_c^{(1)}$ and $\Omega_c^{(2)}$ are the centered kernel matrices calculated as $M_c^{(1)}\Omega^{(1)}M_c^{(1)}$ and $M_c^{(2)}\Omega^{(2)}M_c^{(2)}$ with centering matrix $M_c^{(1)} = I_{n_1} - \frac{1}{n_1}1_{n_1}1_{n_1}^T$ and $M_c^{(2)} = I_{n_2} - \frac{1}{n_2}1_{n_2}1_{n_2}^T$. Here $\Omega^{(1)}$ and $\Omega^{(2)}$ are kernel matrix whose i,j-th elements are calculated as $\Omega_{i,j}^{(1)} = \varphi^{(1)}(x_i^{(1)})^T\varphi^{(1)}(x_j^{(1)})$ and $\Omega_{i,j}^{(2)} = \varphi^{(2)}(x_i^{(2)})^T\varphi^{(2)}(x_j^{(2)})$.

Proof. Consider the Lagrangian of (16),

$$\begin{aligned}\mathcal{L}(w^{(1)}, w^{(2)}, e, r, \alpha, \beta) &= \mu e^T A r - \frac{1}{2}{w^{(1)}}^T w^{(1)} - \frac{1}{2}e^T V_1 e \\ &- \frac{1}{2}{w^{(2)}}^T w^{(2)} - \frac{1}{2}r^T V_2\, r - \alpha^T\Big(e - \Phi_c^{(1)} w^{(1)}\Big) \\ &- \beta^T\Big(r - \Phi_c^{(2)} w^{(2)}\Big),\end{aligned}$$

where α and β are the vector of Lagrange multipliers. Then the Karush-Kuhn-Tucker (KKT) optimality conditions are as follows,

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w^{(1)}} = 0 \rightarrow\ w^{(1)} = {\Phi_c^{(1)}}^T \alpha, \\ \frac{\partial \mathcal{L}}{\partial w^{(2)}} = 0 \rightarrow\ w^{(2)} = {\Phi_c^{(2)}}^T \beta, \\ \frac{\partial \mathcal{L}}{\partial e} = 0 \rightarrow\ \mu A r = V_1 e + \alpha \\ \frac{\partial \mathcal{L}}{\partial r} = 0 \rightarrow\ \mu A^T e = V_2 r + \beta \\ \frac{\partial \mathcal{L}}{\partial \alpha} = 0 \rightarrow\ e = \Phi_c^{(1)} w^{(1)} \\ \frac{\partial \mathcal{L}}{\partial \beta} = 0 \rightarrow\ r = \Phi_c^{(2)} w^{(2)}. \end{cases} \tag{18}$$

Elimination of the primal variables $w^{(1)}, w^{(2)}, e$ and r, results in the generalized eigenvalue problem (17) in the dual. □

B. Explicit Feature Map

The cost of storing and calculating the eigenvectors (17) when n_1 and n_2 are large can be prohibitive. Therefore, next we present an approach that uses an explicit approximate feature map matrices $\hat{\Phi}_c^{(1)}$ and $\hat{\Phi}_c^{(2)}$ and solves the problem in the primal. Thus, the semi-paired kernel canonical correlation analysis with an explicit feature map matrcises $\hat{\Phi}_c^{(1)}$ and $\hat{\Phi}_c^{(2)}$ is formulated in the primal as follows:

$$\begin{aligned}\max_{w^{(1)},w^{(2)},\hat{e},\hat{r}} \quad & \mu e^T A r - \frac{1}{2}{w^{(1)}}^T w^{(1)} - \frac{1}{2}{w^{(2)}}^T w^{(2)} \\ & - \frac{1}{2}\hat{e}^T\Big(\gamma_1 P_1 + (1-\gamma_1)L_1\Big)\hat{e} \\ & - \frac{1}{2}\hat{r}^T\Big(\gamma_2 P_2 + (1-\gamma_2)L_2\Big)\hat{r} \\ \text{subject to} \quad & \hat{e} = \hat{\Phi}_c^{(1)} w^{(1)} \\ & \hat{r} = \hat{\Phi}_c^{(2)} w^{(2)}.\end{aligned} \tag{19}$$

Lemma V.2. *Given a centered finite dimensional approximation to the feature map matrices, $\hat{\Phi}_c^{(1)}$ and $\hat{\Phi}_c^{(2)}$, and regularization constants $\gamma_1, \gamma_2 \in \mathbb{R}^+$, the solution to (19) is obtained by solving the following generalized eigenvalue problem:*

$$\left[\begin{array}{c|c} 0 & (\hat{\Phi}_c^{(1)})^T A \hat{\Phi}_c^{(2)} \\ \hline (\hat{\Phi}_c^{(2)})^T A^T \hat{\Phi}_c^{(1)} & 0 \end{array}\right]\left[\begin{array}{c} w^{(1)} \\ \hline w^{(2)} \end{array}\right] =$$

$$\lambda \left[\begin{array}{c|c} \frac{(\hat{\Phi}_c^{(1)})^T V_1 (\hat{\Phi}_c^{(1)}) + I_{n_1}}{0} & \frac{0}{(\hat{\Phi}_c^{(2)})^T V_2 (\hat{\Phi}_c^{(2)}) + I_{n_2}} \end{array}\right] \left[\frac{w^{(1)}}{w^{(2)}}\right], \quad \text{and} \tag{20}$$

where V_1 and V_2 are defined as previously in (17).

Proof. Since an explicit feature map matrices $\hat{\Phi}_c^{(1)}$ and $\hat{\Phi}_c^{(2)}$ are known, one can eliminate the primal variables $\hat{r}$ and $\hat{e}$ and reformulate (19) as an unconstrained optimization problem. Setting the derivatives of the cost function with respect to the primal variables $w^{(1)}$ and $w^{(2)}$ to zero, yields the generalized eigenvalue problem (20). □

Remark V.1. *In the formulation (16), where an implicit feature map is used, one solves a generalized eigenvalue problem of size $(n_1+n_2)\times(n_1+n_2)$ in the dual. Therefore the algorithm has $\mathcal{O}((n_1+n_2)^3)$ training complexity with naive implementation. On the other hand, when an explicit feature map is used one can solve the problem in the primal. Given $N = n_1 + n_2$ training points, the complexity of calculating the Nyström approximation, with eigenvalue decomposition of the kernel matrix of size m and solving a generalized eigenvalue problem of size $m \times m$ are $\mathcal{O}(m^3 + Nm^2)$ and $\mathcal{O}(m^3)$ respectively where $m = m_1 + m_2$ and $m \ll N$. The total complexity of the proposed method, neglecting lower order terms, is then given by the sum of the two complexities.*

VI. Regularized Semi-Paired KCCA

The approaches discussed in the previous section, are unsupervised techniques i.e. they did not use the label information of the data points. However, as collecting paired data points, it is also costly to get labeled data points. Therefore in practice, one may have access to few labeled data points from either the paired data or unpaired ones as well as large amount of unlabeled data. In this situation, using a semi-supervised learning algorithm that can learn from both labeled and unlabeled data is preferable [33], [34], [37], [38]. Sun et al. [39] proposed the discriminative canonical correlation analysis (DCCA) for supervised multi-modality data. DCCA maximizes the within-class correlation while minimizing the between-class correlation. Here we assume that the small amount of labeled and unlabeled paired data are available. In addition in our formulation we also consider integration of the unpaired data from both domains. Therefore, we aim at incorporating the information of the labeled, unlabeled, paired and unpaired data points of the source and target domains in the learning process through a unif ed optimization problem in order to learn a new representation of the data that have more discriminative power. All the available information are encapsulated in the primal optimization problem using appropriate regularization terms.

Consider two training datasets $\mathcal{D}^{(1)}$ and $\mathcal{D}^{(2)}$ as source and target domains with

$$\mathcal{D}^{(1)} = \{ \underbrace{x_1^{(1)}, ..., x_{n_p}^{(1)}}_{\substack{paired \\ (labeled/unlabeled) \\ (\mathcal{D}_{p,ul}^{(1)} \cup \mathcal{D}_{p,l}^{(1)})}} , \underbrace{x_{n_p+1}^{(1)}, .., x_{n_1}^{(1)}}_{\substack{unpaired \\ (unlabeled) \\ (\mathcal{D}_{up,ul}^{(1)})}} , \underbrace{x_{n_p+1}^{(1)}, .., x_{N_1}^{(1)}}_{\substack{unpaired \\ (labeled) \\ (\mathcal{D}_{up,l}^{(1)})}} \},$$

$$\mathcal{D}^{(2)} = \{ \underbrace{x_1^{(2)}, ..., x_{n_p}^{(2)}}_{\substack{paired \\ (labeled/unlabeled) \\ (\mathcal{D}_{p,ul}^{(2)} \cup \mathcal{D}_{p,l}^{(2)})}} , \underbrace{x_{n_p+1}^{(2)}, .., x_{n_2}^{(2)}}_{\substack{unpaired \\ (unlabeled) \\ (\mathcal{D}_{up,ul}^{(2)})}} \},$$

where $\mathcal{D}_p^{(1)}$ and $\mathcal{D}_p^{(2)}$ are paired samples whereas $\mathcal{D}_{up}^{(1)}$ and $\mathcal{D}_{up}^{(2)}$ are unpaired samples.

Often in domain adaptation setting, the number of labeled instances in the target domain is limited, compared to that of the source domain. Therefore, here we assume that only a small number of paired labeled data points from both domains are available, i.e. $\{x_1^{(i)}, \ldots, x_{n_L}^{(i)}\}$ from $\mathcal{D}_p^{(1)}$ and $\mathcal{D}_p^{(2)}$ are labeled for $i = 1, 2$. Furthermore, the source dataset is equipped with additional unpaired labeled instances during training. Assuming that there are Q classes, we def ne the label indicator matrix $Y^{(1)} \in \mathbb{R}^{n_{L_1} \times Q}$ for the source domain as follows:

$$Y_{ij}^{(1)} = \begin{cases} +1 & \text{if the } i\text{-th point belongs to the } j\text{-th class} \\ -1 & \text{otherwise,} \end{cases} \tag{21}$$

where the total number of labeled instances in the source domain is denoted by n_{L_1}. Similarly one can def ne the label indicator matrix $Y^{(2)} \in \mathbb{R}^{n_{L_2} \times Q}$ for the target domain where n_{L_2} denotes the total number of labeled instances in the target domain.

A. Implicit Feature Map

The Regularized Semi-Paired KCCA (RSP-KCCA) with centered implicit feature map matrices $\Phi_c^{(1)}$ and $\Phi_c^{(2)}$ in the primal is formulated as follows:

$$\begin{aligned} \max_{w_\ell^{(1)}, w_\ell^{(2)}, r_\ell, e_\ell} \quad & \mu \sum_{\ell=1}^{Q} e_\ell^T A r_\ell - \frac{1}{2}\sum_{\ell=1}^{Q} {w_\ell^{(1)}}^T w_\ell^{(1)} - \frac{1}{2}\sum_{\ell=1}^{Q} {w_\ell^{(2)}}^T w_\ell^{(2)} \\ & - \frac{1}{2}\sum_{\ell=1}^{Q} {e_\ell}^T V_1 e_\ell - \frac{1}{2}\sum_{\ell=1}^{Q} {r_\ell}^T V_2 r_\ell \\ & + \frac{\gamma_3}{2}\sum_{\ell=1}^{Q} e_\ell^T c_\ell^{(1)} + r_\ell^T c_\ell^{(2)} \\ \text{subject to} \quad & e_\ell = \Phi_c^{(1)} w_\ell^{(1)}, \ \ell = 1, \ldots, Q, \\ & r_\ell = \Phi_c^{(2)} w_\ell^{(2)}, \ \ell = 1, \ldots, Q, \end{aligned} \tag{22}$$

where $c_\ell^{(1)}$ is the ℓ-th column of the matrix $C^{(1)}$ def ned as

$$C^{(1)} = [c_1^{(1)}, \ldots, c_Q^{(1)}]_{n_1 \times Q} = \left[\frac{Y^{(1)}}{0_{n_{u_1} \times Q}}\right]_{N_1 \times Q}, \tag{23}$$

where the subscript n_{u_1} denotes the total number of unlabeled instances from source domain. $0_{n_{u_1} \times Q}$ is a zero matrix of size $n_{u_1} \times Q$ and $Y^{(1)}$ is def ned as previously. Similarly $c_\ell^{(2)}$ is the ℓ-th column of the matrix $C^{(2)}$ def ned as

$$C^{(2)} = [c_1^{(2)}, \ldots, c_Q^{(2)}]_{n_2 \times Q} = \left[\frac{Y^{(2)}}{0_{n_{u_2} \times Q}}\right]_{n_2 \times Q}, \tag{24}$$

where the subscript n_{u_2} denotes the total number of unlabeled instances from target domain. Here matrix A as follows,

$$A = \left[\begin{array}{c|c} I_{n_p \times n_p} & 0_{n_p \times n_2 - n_p} \\ \hline 0_{N_1 - n_p \times n_p} & 0_{N_1 - n_p \times n_2 - n_p} \end{array}\right]_{N_1 \times n_2},$$

$V_1 = \gamma_1 P_1 + (1-\gamma_1)L_1$ and $V_2 = \gamma_2 P_2 + (1-\gamma_2)L_2$ where P_2 is def ned as previously and P_1 is def ned as follows:

$$P_1 = \left[\begin{array}{c|c} I_{n_p \times n_p} & 0_{n_p \times N_1 - n_p} \\ \hline 0_{N_1 - n_p \times n_p} & 0_{N_1 - n_p \times N_1 - n_p} \end{array}\right]_{N_1 \times N_1}.$$

Lemma VI.1. *Given two positive definite kernel functions $K_1(s,t) = \varphi^{(1)}(s)^T \varphi^{(1)}(t)$, $K_2(s,t) = \varphi^{(2)}(s)^T \varphi^{(2)}(t)$ and regularization constants γ_1, γ_2, γ_3 and μ, the solution to the optimization problem (22) is obtained by solving the following linear system of equations,*

$$\left[\begin{array}{c|c} V_1\Omega_c^{(1)} + I_{N_1} & -\mu A \Omega_c^{(2)} \\ \hline -\mu A^T \Omega_c^{(1)} & V_2 \Omega_c^{(2)} + I_{n_2} \end{array}\right] \left[\begin{array}{c} \alpha_\ell \\ \hline \beta_\ell \end{array}\right] = \gamma_3 \left[\begin{array}{c} c_\ell^{(1)} \\ \hline c_\ell^{(2)} \end{array}\right], \ \ell = 1, \ldots, Q, \tag{25}$$

where α_ℓ, β_ℓ are Lagrange multiplier vectors. $\Omega_c^{(1)}$ and $\Omega_c^{(2)}$ are the centered kernel matrices. calculated as $M_c^{(1)}\Omega^{(1)}M_c^{(1)}$ and $M_c^{(2)}\Omega^{(2)}M_c^{(2)}$ with centering matrix $M_c^{(1)} = I_{N_1} - \frac{1}{N_1}1_{N_1}1_{N_1}^T$ and $M_c^{(2)} = I_{n_2} - \frac{1}{n_2}1_{n_2}1_{n_2}^T$. Here $\Omega^{(1)}$ and $\Omega^{(2)}$ are kernel matrix whose i,j-th elements are calculated as $\Omega_{i,j}^{(1)} = \varphi^{(1)}(x_i^{(1)})^T \varphi^{(1)}(x_j^{(1)})$ and $\Omega_{i,j}^{(2)} = \varphi^{(2)}(x_i^{(2)})^T \varphi^{(2)}(x_j^{(2)})$.

Proof. The Lagrangian of the constrained optimization problem (22) becomes

$$\begin{aligned}
\mathcal{L}(w_\ell^{(1)}, w_\ell^{(2)}, e_\ell, r_\ell, \alpha_\ell, \beta_\ell) &= \mu \sum_{\ell=1}^{Q} e_\ell^T A r_\ell - \frac{1}{2}\sum_{\ell=1}^{Q} w_\ell^{(1)^T} w_\ell^{(1)} \\
&- \frac{1}{2}\sum_{\ell=1}^{Q} w_\ell^{(2)^T} w_\ell^{(2)} - \frac{1}{2}\sum_{\ell=1}^{Q} e_\ell{}^T V_1 e_\ell - \frac{1}{2}\sum_{\ell=1}^{Q} r_\ell{}^T V_2 r_\ell \\
&+ \frac{\gamma_3}{2}\sum_{\ell=1}^{Q} e_\ell^T c_\ell^{(1)} + r_\ell^T c_\ell^{(2)} - \sum_{\ell=1}^{Q} \alpha_\ell{}^T \Big(e_\ell - \Phi_c^{(1)} w_\ell^{(1)}\Big), \\
&- \sum_{\ell=1}^{Q} \beta_\ell{}^T \Big(r_\ell - \Phi_c^{(2)} w_\ell^{(2)}\Big),
\end{aligned}$$

where α_ℓ and β_ℓ are the Lagrange multiplier vectors. Then the Karush-Kuhn-Tucker (KKT) optimality conditions are as follows,

$$\begin{cases}
\frac{\partial \mathcal{L}}{\partial w_\ell^{(1)}} = 0 \rightarrow w_\ell^{(1)} = \Phi_c^{(1)^T} \alpha_\ell, \ell = 1, \ldots, Q, \\
\frac{\partial \mathcal{L}}{\partial w_\ell^{(2)}} = 0 \rightarrow w_\ell^{(2)} = \Phi_c^{(2)^T} \beta_\ell, \ell = 1, \ldots, Q, \\
\frac{\partial \mathcal{L}}{\partial e_\ell} = 0 \rightarrow \mu A r_\ell = V_1 e_\ell + \alpha_\ell + \gamma_3 c_\ell^{(1)}, \ell = 1, \ldots, Q, \\
\frac{\partial \mathcal{L}}{\partial r_\ell} = 0 \rightarrow \mu A^T e_\ell = V_2 r_\ell + \beta_\ell + \gamma_3 c_\ell^{(2)}, \ell = 1, \ldots, Q, \\
\frac{\partial \mathcal{L}}{\partial \alpha_\ell} = 0 \rightarrow e_\ell = \Phi_c^{(1)} w_\ell^{(1)}, \ell = 1, \ldots, Q, \\
\frac{\partial \mathcal{L}}{\partial \beta_\ell} = 0 \rightarrow r_\ell = \Phi_c^{(2)} w_\ell^{(2)}, \ell = 1, \ldots, Q,
\end{cases} \tag{26}$$

Eliminating the primal variables $w_\ell^{(1)}, w_\ell^{(2)}, e_\ell, r_\ell$, for $\ell = 1, \ldots, Q$, results in the linear system of equations (25) in the dual. □

The score variables for the training dataset of the source and target domains can be expressed as follows:

$$\begin{cases} z_e^\ell = \Phi_c^{(1)} w_\ell^{(1)} = \Omega_c^{(1)} \alpha_\ell \ \ell = 1, \ldots, Q, \\ z_r^\ell = \Phi_c^{(2)} w_\ell^{(2)} = \Omega_c^{(2)} \beta_\ell \ \ell = 1, \ldots, Q. \end{cases} \tag{27}$$

Similarly, one can compute the score variables of the new unseen test data points by projecting the centered explicit feature map of the test points onto the learned solution vectors $w^{(1)}$ and $w^{(2)}$.

In Equation (22) the feature map is not explicitly known, therefore one uses the kernel trick and solves a linear system of size $(N_1+n_2)\times(N_1+n_2)$ (number of data points) in the dual. For large scale data, it is not appropriate to solve the problem in the dual. In what follows we show how one can use the approximation of the feature map (explained in section III) to solve the problem in the primal.

B. Explicit Feature Map

Consider the problem (22), now with explicit approximate of the feature map matrices $\hat{\Phi}_c^{(1)}$ and $\hat{\Phi}_c^{(2)}$ obtained by the Nyström Method (see section III). Then the following lemma can be obtained.

Lemma VI.2. *Given a centered finite dimensional approximation to the feature map matrices, $\hat{\Phi}_c^{(1)}$ and $\hat{\Phi}_c^{(2)}$, and regularization constants $\mu, \gamma_1, \gamma_2, \gamma_3 \in \mathbb{R}^+$, the solution to (22) is obtained by solving the following linear system of equations:*

$$\left[\begin{array}{c|c} (\hat{\Phi}_c^{(1)})^T V_1 \hat{\Phi}_c^{(1)} + I_{m_1} & -\mu (\hat{\Phi}_c^{(1)})^T A \hat{\Phi}_c^{(2)} \\ \hline -\mu (\hat{\Phi}_c^{(2)})^T A^T \hat{\Phi}_c^{(1)} & (\hat{\Phi}_c^{(2)})^T V_2 \hat{\Phi}_c^{(2)} + I_{m_2} \end{array}\right] \left[\begin{array}{c} w_\ell^{(1)} \\ \hline w_\ell^{(2)} \end{array}\right] = \gamma_3 \left[\begin{array}{c} (\hat{\Phi}_c^{(1)})^T c_\ell^{(1)} \\ \hline (\hat{\Phi}_c^{(2)})^T c_\ell^{(2)} \end{array}\right], \ \ell = 1, \ldots, Q, \tag{28}$$

where V_1 and V_2 are defined as previously.

Proof. Given the explicit feature maps, one can rewrite (22) as an unconstrained optimization problem.

Setting the derivatives of the cost function (in the obtained unconstrained optimization problem) with respect to the primal variables $w^{(1)}$ and $w^{(2)}$ to zero, yields,

$$\begin{cases}
\frac{\partial \mathcal{J}}{\partial w_\ell^{(1)}} = 0 \Rightarrow \\
\Big[(\hat{\Phi}_c^{(1)})^T V_1 \hat{\Phi}_c^{(1)} + I_{m_1}\Big] w_\ell^{(1)} - \mu (\hat{\Phi}_c^{(1)})^T A \hat{\Phi}_c^{(2)} w_\ell^{(2)} + = \\
\qquad \gamma_3 \hat{\Phi}_c^{(1)T} c_\ell^{(1)}, \ \ell = 1, \ldots, Q \\
\frac{\partial \mathcal{L}}{\partial w_\ell^{(2)}} = 0 \Rightarrow \\
-\mu (\hat{\Phi}_c^{(2)})^T A^T \hat{\Phi}_c^{(1)} w_\ell^{(1)} + \Big[(\hat{\Phi}_c^{(2)})^T V_2 \hat{\Phi}_c^{(2)} + I_{m_2}\Big] w_\ell^{(2)} = \\
\qquad \gamma_3 (\hat{\Phi}_c^{(2)})^T c_\ell^{(2)}, \ \ell = 1, \ldots, Q
\end{cases} \tag{29}$$

where with some algebraic manipulation one can rewrite the above equation as linear system of equations in (28).

□

The score variables for training data can be expressed as follows:

$$\begin{cases} \hat{z}_e^{\ell} = \hat{\Phi}_c^{(1)} w_{\ell}^{(1)} \ \ell = 1, \ldots, Q \\ \hat{z}_r^{\ell} = \hat{\Phi}_c^{(2)} w_{\ell}^{(2)} \ \ell = 1, \ldots, Q. \end{cases} \tag{30}$$

Similarly, one can compute the score variables of the new unseen test data points by projecting the centered explicit feature map of the test points onto the learned solution vectors $w^{(1)}$ and $w^{(2)}$.

Remark VI.1. *It should be noted that by incorporating the labels into the primal Semi-Paired KCCA formulation (16), the dual formulation is changed from a generalized eigenvalue problem (17) to a system of linear equations (25). Assuming that the given dataset has Q classes, due the fact that a linear system is solved in (25), the number of solution vectors is Q. Therefore the source and target domains $\mathcal{D}^{(1)}$ and $\mathcal{D}^{(2)}$ will be projected into a Q-dimensional space as opposed to solution of the eigenvalue problem (17), where the data points can be projected into a $n_1 + n_2$-dimensional space.*

Remark VI.2. *In the formulation (22), where an implicit feature map is used, one solves a linear system of equations of size $(N_1 + n_2) \times (N_1 + n_2)$ in the dual. Therefore these algorithms have $\mathcal{O}((N_1 + n_2)^3)$ training complexity with naive implementation. On the other hand, when an explicit feature map is used one can solve the problem in the primal. Given $N = N_1 + n_2$ training points, The complexity of calculating the Nyström approximation, with eigenvalue decomposition of the kernel matrix of size m and solving a linear system of equations of size $m \times m$ are is $\mathcal{O}(m^3 + Nm^2)$ and $\mathcal{O}(m^3)$ respectively where $m = m_1 + m_2$ and $m \ll N$. The total complexity of the proposed method, neglecting lower order terms, is then given by the sum of the two complexities.*

The general two step procedure for learning a domain-invariant features followed by the semi-supervised classif er, MSS-KSC [34] is described in Fig. 3. MSS-KSC uses a purely unsupervised algorithm, Kernel Spectral Clustering, as a core model and the labels are integrated into the model in order to guide the clustering result.

Table 1, summarizes the type of instances that can be used by the proposed approaches in order to learn new representation of the source and target instances.

TABLE I
TYPE OF INSTANCES USED BY THE PROPOSED MODELS

	Method		
Data points	KCCA	Semi-Paired-KCCA	Regularized-Semi-Paired-KCCA
(paired, unlabeled)	✓	✓	✓
(unpaired, unlabeled)	✗	✓	✓
(paired, labeled)	✗	✗	✓
(unpaired, labeled)	✗	✗	✓

The proposed approach is summarized in Algorithm 1.

Algorithm 1: Proposed two-step approach

Input: Training data set $\mathcal{D}^{(1)}$ and $\mathcal{D}^{(2)}$ from both source and target domains, test data points $\mathcal{D}_{test}$ (from both source and target domains), the tuning parameters $\{\gamma_i\}_{i=1}^3$, μ and the kernel parameter (if any), number of available class labels i.e. Q

Output: Class membership of test data points.

First stage: LEARNING DOMAIN INVARIANT FEATURES.

Use the training data set $\mathcal{D}^{(1)}$ and $\mathcal{D}^{(2)}$ (from both source and target domains).

Learn new representation of training data, $\mathcal{D}^{(1)}$ and $\mathcal{D}^{(2)}$ by solving:

- (i) Equation (13) in the case of unlabeled paired cross data points (see Fig. 2(a)).
- (ii) Equation (20) in the case of unlabeled semi-paired cross data points (see Fig. 2(b)).
- (iii) Equation (28) in the case of Unlabeled and partially labeled paired cross-domain data (see Fig. 2(c)).

Use the out-of-sample extension of the proposed models to estimate the new representation of test data points of both domains.

Second stage: PREDICT THE LABELS OF THE TEST TARGET DOMAIN DATA

Train the MSS-KSC model [34], using the obtained representation of the training data points. [1]

predict the labels of the test target data points using the learned MSS-KSC model.

VII. MODEL SELECTION

The performance of the proposed methods depends on the choice of the tuning parameters. In this paper for all the experiments the Gaussian RBF kernel is used. The optimal values of the regularization constants γ_1, γ_2, γ_3, μ and the kernel bandwidth parameter σ are obtained by evaluating the performance of the model on the validation set. For the KCCA and Semi-Paired KCCA models, the optimal model parameters can be selected such that the correlation coeff cients of the score variables $z_e^{(i)}$ and $z_r^{(i)}$ is maximized

$$\max_{i \in \{1,\ldots,m_1+m_2\}} \frac{{z_e^{(i)}}^T z_r^{(i)}}{\|z_e^{(i)}\|_2 \|z_r^{(i)}\|_2}.$$

For the Regularized Semi-Paired KCCA approach, since both labeled and unlabeled data points are used in the formulation, the solution vectors preserve the label predictive information. Therefore, once the solution vectors α_ℓ and β_ℓ are obtained for $\ell = 1, \ldots, Q$, one can start making prediction by decoding the score variables of the unseen points obtained using the out-of-sample extension. Therefore the model selection follows a combination of maximizing the correlation coeff cients between

[1] MSS-KSC can learn from both labeled and unlabeled data points. Here the MSS-KSC approach is trained using the projected unlabeled and labeled source data.

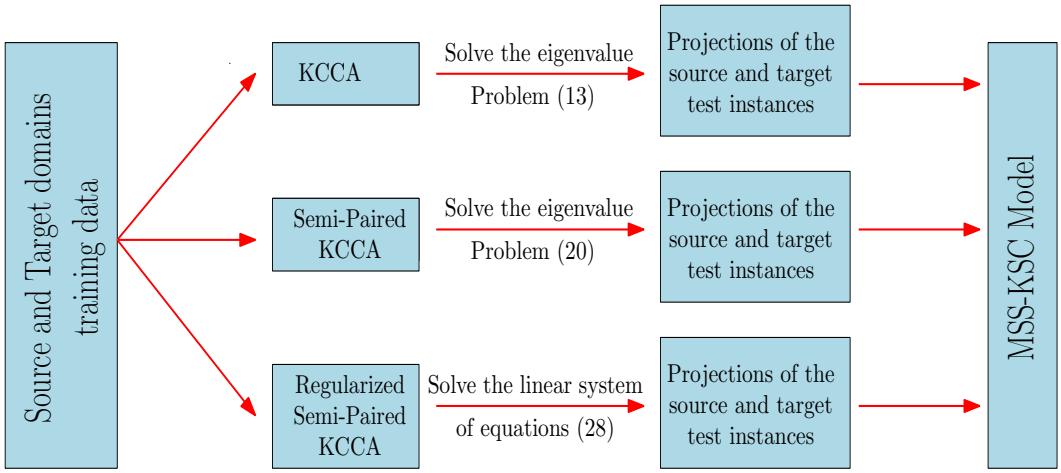


Fig. 3. Flow-chart of the general procedure for learning domain-invariant features for both source and target domains followed by the semi-supervised classifier MSS-KSC [34]. The MSS-KSC model is learned using the projections of the unlabeled and labeled instances of the source domain and the membership of the instances in the target domain is estimated by means of the out-of-sample extension.

the score variables $z_e^{(\ell)}$ and $z_r^{(\ell)}$ as well maximizing the classification accuracy on the validation set, i.e.

$$\underset{\gamma_1,\gamma_2,\gamma_3,\sigma}{\operatorname{argmax}} \ \kappa Corr(\gamma_1,\gamma_2,\gamma_3,\sigma) + (1-\kappa)Acc(\gamma_1,\gamma_2,\gamma_3,\sigma),$$

and $\kappa \in [0,1]$.

Remark VII.1. *In general, in the proposed KCCA, SP-KCCA and RSP-KCCA models one can use two different feature maps $\varphi^{(1)}$ and $\varphi^{(2)}$ corresponding to source and target domains which result in tuning two kernel parameters in the dual. Here in our simulations, for the sake of simplicity, the kernel bandwidth σ is chosen to be the same for both domains when constructing the kernel matrix in the dual.*

Remark VII.2. *The solution vectors obtained by the Regularized Semi-Paired KCCA approach preserve the label predictive information. Therefore, in RSP-KCCA model, the out-of-sample extension can be performed by decoding the score variables of the unseen points thus avoiding applying additional models for making prediction.*

VIII. NUMERICAL EXPERIMENTS

In this section experimental results on one synthetic as well as four real-life datasets, i.e. Multiple Features dataset, Wikipedia dataset, Office benchmark domain adaptation dataset and the IXMAS action dataset are reported. The experiments are performed on a laptop computer with Intel Core i7 CPU and 8 GB RAM under Matlab 2015a.

In the Wikipedia dataset [40] each instance is associated with a text-image pair. In this dataset, the text component is represented as a sample from a hidden topic model, learned with latent Dirichlet allocation, and images are represented as bags of visual (SIFT) features. In our analysis, five subject categories i.e. art and architecture, biology, literature, sport, and warfare are considered. From each category 100 instances is chosen in our experiments. So the total number of instances are 500.

The Multiple Features Dataset is taken from the UCI Machine Learning repository [41]. It contains ten classes, $(0-9)$, of handwritten digits with 200 images per class, thus for a total of 2000 images. These digits are represented in terms of six different types of features (heterogeneous features) with different dimensionalities. The features are Fourier coefficients of the character shapes (fou: 76), profile correlations (fac: 216), Karhunen-Loeve coefficients (kar: 64), pixel average (pix: 240), Zernike moments (zer: 47) and morphological features (mor: 6). The abbreviation as well as the number of dimension of each feature is given in the parenthesis.

The Office benchmark domain adaptation dataset [6], [42] offers two domains 'webcam' and 'dslr'. The webcam domain is a collection of objects in an office environment taken with a webcam. Similarly, the dslr domain is a collection of the same objects taken with a DSLR camera. Therefore, there is a resolution and lighting domain shift. There are in total 31 classes. Here we have extracted 10 classes Backpack, Touring-bike, Calculator, Headphones, Computer-Keyboard, Laptop-101, Computer-Monitor, Computer-Mouse, Coffee mug, and Video-Projector from both domains. Fig. 4 highlights the differences among the two domains with example images from the category of Monitor and Computer-Keyboard. As introduced in [6], all images are represented by the SURF bag of words feature vectors quantized to 800 dimensions. Following [6], we first apply PCA to the data of both domains to lower the dimensionality of the data to 30.

The fourth dataset is the IXMAS action dataset which has been previously used to address cross-view action recognition problems in [43], [44]. This dataset contains videos of eleven action classes. Each action video is performed three times by twelve actors forming 36 instances per class. The actions are

simultaneously captured by five cameras which provide five domains, see Fig. 5. Following the feature extraction settings as in [43] the same type of features of size 1000 in both source and target domains are extracted.

We consider the cross-domain classification task as a domain adaptation problem. The source domain data obtained from one of the feature spaces is used for training the model, while the target domain which represents another feature space are used as the test data. We have conducted a comparison with five representative and state-of-the-art domain adaptation algorithms, HFA, GFK, mSDA, TCA and MIDA proposed in [45], [3], [46], [7] and [47] respectively. The Heterogeneous Feature Augmentation (HFA) algorithm proposed in [45] uses two different projection matrices to transform the labeled data from two heterogeneous domains into a common subspace followed by two new feature mapping functions to augment the transformed data with their original features and zeros.

The Geodesic Flow Kernel (GFK) approach proposed in [3] models domain shift by integrating an infinite number of subspaces that characterize changes in geometric and statistical properties from the source to the target domain. The marginalized stacked denoising autoencoders (mSDA) is proposed in [46] to learn new representations for domain adaptation. As opposed to SDAs, mSDA does not require stochastic gradient descent or other optimization algorithms to learn the parameters and that there is a closed-form solution for obtaining the model parameters. The authors showed that the representations learned by mSDA are as effective as the traditional SDAs in benchmarked tasks. As suggested in [46], both domains not necessarily should use identical features and one can pad all input vectors with zeros to make both domains be of equal dimensionality.

Transfer component analysis (TCA) introduced in [7] learns transfer components across domains using maximum mean discrepancy. It is also further extended to semi-supervised TCA (SSTCA) to leverage label information and preserve local geometry of the manifold. The semi-supervised MIDA approach proposed in [47] learns a subspace which has maximum independence with the domain features, so as to reduce the inter-domain discrepancy in distributions.

In all the experiments the noise level parameter in mSDA approach has been tuned in the range $[10^{-2}, 10^{0}]$ and the number of layers is set to 3. The default parameters for the Heterogeneous Feature Augmentation, Geodesic Flow Kernel, TCA and MIDA approaches are used.

The following experimental setup is used in this paper. We first select at random 50% from both soucre and target data in order to construct the paired training samples i.e. $\mathcal{D}_p^{(1)}$ and $\mathcal{D}_p^{(2)}$. Then 20% of the remaining data from both source and target domains is selected at random to be used as unpaired and unlabeled samples $\mathcal{D}_{up,ul}^{(1)}$ and $\mathcal{D}_{up,ul}^{(2)}$. The final remaining 30% of the source data are labeled in order to construct $\mathcal{D}_{up,l}^{(1)}$ (unpaired and unlabeled). The same remaning percentage of the target domain data defines the test target dataset, $\mathcal{D}_{\text{test}}^{(2)}$.

In many real life applications obtaining labels for target domain data is costly. Therefore here we only assign labels to two randomly selected samples per class from $\mathcal{D}_p^{(1)}$ and $\mathcal{D}_p^{(2)}$ in order to construct $\mathcal{D}_{p,l}^{(1)}$ and $\mathcal{D}_{p,l}^{(2)}$ sets. The remaining data points of $\mathcal{D}_p^{(1)}$ and $\mathcal{D}_p^{(2)}$ are unlabeled and construct $\mathcal{D}_{p,ul}^{(1)}$ and $\mathcal{D}_{p,ul}^{(2)}$ sets.

The KCCA approach is trained using paired and unlabeled instances, i.e. $\mathcal{D}^{(1)} = \mathcal{D}_{p,ul}^{(1)}$ and $\mathcal{D}^{(2)} = \mathcal{D}_{p,ul}^{(2)}$. For the SP-KCCA model, we include unpaired unlabeled instances from both domains into the training set, i.e. $\mathcal{D}^{(1)} = \mathcal{D}_{p,ul}^{(1)} \cup \mathcal{D}_{up,ul}^{(1)}$ and $\mathcal{D}^{(2)} = \mathcal{D}_{p,ul}^{(2)} \cup \mathcal{D}_{up,ul}^{(2)}$. Finally for the RSP-KCCA approach we also provide paired labeled data from source and target domains in the training stage. In addition, the training source domain data also includes unpaired labeled instances. Therefore, RSP-KCCA model is trained using $\mathcal{D}^{(1)} = \mathcal{D}_{p,ul}^{(1)} \cup \mathcal{D}_{p,l}^{(1)} \cup \mathcal{D}_{up,ul}^{(1)} \cup \mathcal{D}_{up,l}^{(1)}$ and $\mathcal{D}^{(2)} = \mathcal{D}_{p,ul}^{(2)} \cup \mathcal{D}_{p,l}^{(2)} \cup \mathcal{D}_{up,ul}^{(2)}$, (see Table I). In all the cases, $\mathcal{D}_{\text{test}}^{(2)}$ is used as test target domain to evaluate the performance of the algorithms. Furthermore, the accuracy of proposed models on the unlabeled source data pints is also reported.

The two proposed models KCCA, SP-KCCA learn the domain-invariant features for the source and target data points using unlabeled source and target domain datasets. The RSP-KCCA approach learns the domain-invariant features for the source and target data points using labeled and unlabeled source and target domain datasets (the number of labeled target domains is limited). As opposed to KCCA and SP-KCCA models, the solution vectors obtained by RSP-KCCA model preserve the label predictive information. Therefore, once the solution vectors α_ℓ and β_ℓ are obtained for $\ell = 1, \ldots, Q$, one can start making prediction by decoding the score variables of the unseen points obtained using the out-of-sample extension property of the model.

The performance of the proposed RSP-KCCA approach is illustrated on a synthetic example. Consider the two moons shown in Fig. 6a, one as source domain and the other as target domain. In this case, the target domain is data points of the source domain rotated and shifted. The learned features by the proposed approach are depicted in Fig. 6. From Fig 6, one can observe that there is a very good agreement between the projected variables of both domains, i.e. e_1, r_1 and e_2, r_2. In addition, thanks to the regularization mechanism, used in the formulation of RSP-KCCA, the projected variables are encouraged to be as close as possible to the labels provided by the user. Therefore the learned features are endowed with a discriminative power.

The results of the proposed models as well as those obtained by Heterogeneous Feature Augmentation (HFA) algorithm proposed in [45] and marginalized stacked denoising autoencoders (mSDA) [46] on Wikipedia dataset is shown in Fig. 7. Notice that as the dimension of the two domains are not the same, therefore only algorithms that can address heterogeneous features are employed. For the mSDA approach, as the authors in [46] suggested we have pad the input vectors with zeros to make both domains be of equal dimension. One can observe that the learned features using RSP-KCCA approach have slightly better discriminative power than those obtained by other models. The accuracy of the SVM classifier trained on the original and not transferred source domain data and tested on the target domain data, under the same setting as the other methods, are as follows: $image - text$: 0.86 and $text - image$: 0.25. From Fig. 7, it is



Fig. 4. **First row:** Images of keyboard and monitors from webcam domain. **Second row:** Images of keyboard and monitors from dslr domain.

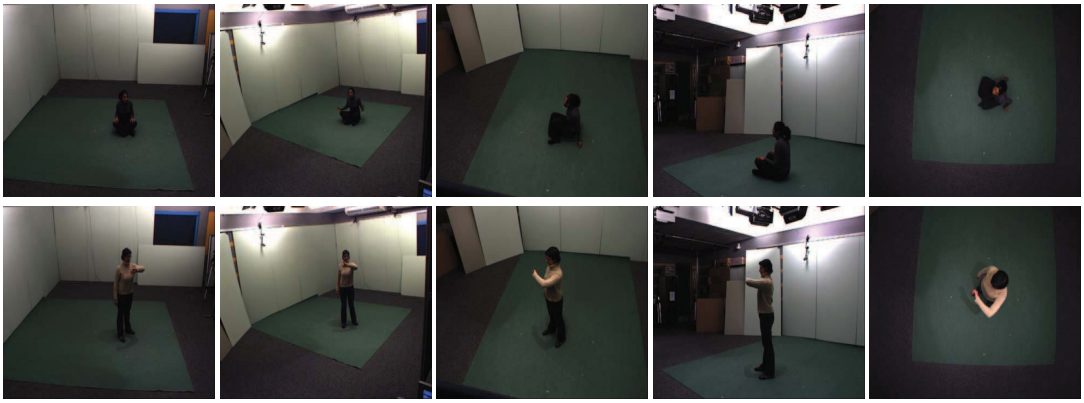
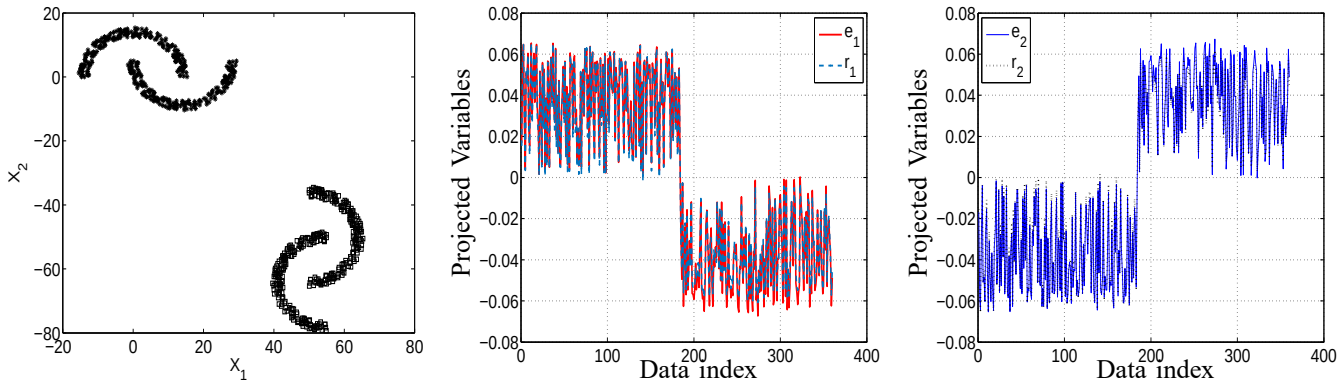


Fig. 5. Two example actions of the IXMAS dataset. **First row:** Represents 'sit down' action at five different domains. **Second row:** Represents 'check watch' action at five different domains.



(a) (b) (c)

Fig. 6. Learned domain-invariant features using RSP-KCCA model. **(a)** Two-moons datasets. The left two-moons are source domains data and the right two-moons are considered to be target data. **(b)** Projected variables e_1 and r_1 for the source domain (left two-moons) and target domain (right two-moons). **(c)** Projected variables e_2 and r_2 for the source domain (left two-moons) and target domain (right two-moons).

apparent that applying domain adaptation techniques enhances the performance of the classifier for this dataset.

Fig. 8 shows the test accuracy of different models on the Office dataset. The RSP-KCCA model outperforms the other approaches in source-to-source cases while has a comparable results with respect to other approaches in the source-target cases. The feature dimension of both domains for this dataset is the same.

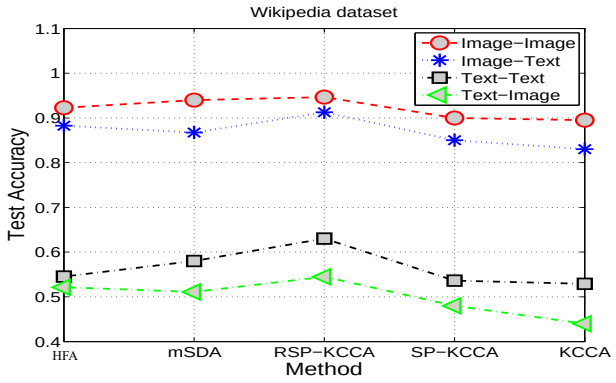


Fig. 7. The average test accuracy of five methods HFA [45], mDSA [46], RSP-KCCA, SP-KCCA and KCCA on source and target domains for Wikipedia dataset over 10 simulation runs.

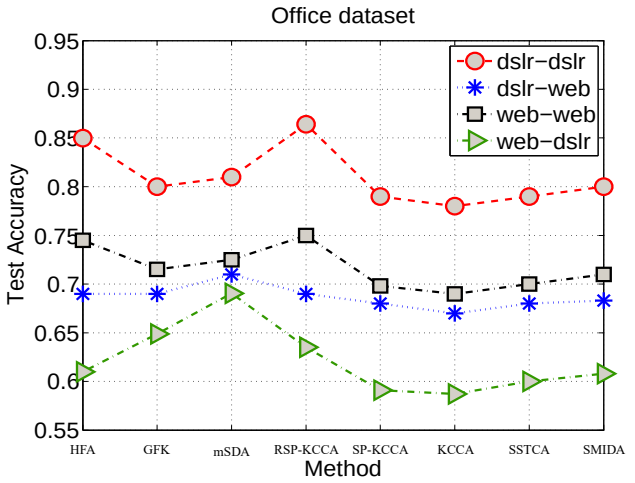


Fig. 8. The average test accuracy of eight methods HFA [45], GFK [3], mSDA [46], RSP-KCCA, SP-KCCA, KCCA, SSTCA and SMIDA on the source and target domains for office dataset over 10 simulation runs.

It should be mentioned that the linear correlation between the projected training data from source and target domains does not exhibit the same pattern for three KCCA, Semi-paired KCCA and Regularized Semi-Paired KCCA models. In particular, from Fig. 9 one may notice that the correlation between the learned features using Regularized Semi-Paired KCCA exhibits a different structure than those obtained by Semi-Paired KCCA and KCCA models. This is due to the fact that the Regularized Semi-Paired KCCA model do carry the labels information. The average accuracy of different models on Multiple features and IXMAS action datasets over 10 simulation runs are tabulated in Table II and III. One can observe that the features learned by Regularized Semi-Paired KCCA most often leads to a better classification accuracy than the other approaches for Multiple features datasets. The result of RSP-KCCA approach on IXMAS action dataset is comparable to that of HFA model and better than those of the other models.

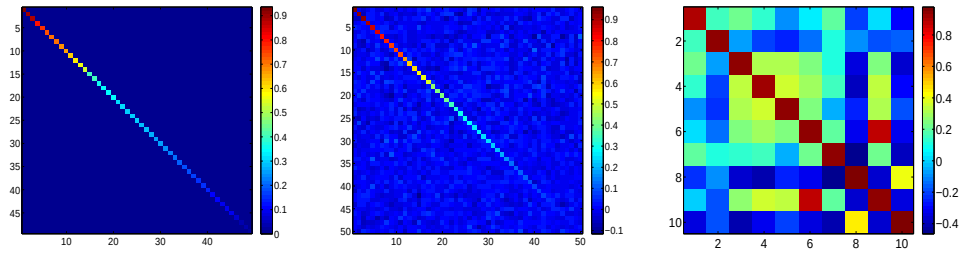
TABLE II
COMPARING THE AVERAGE TEST ACCURACY OF THE PROPOSED RSP-KCCA MODEL WITH THOSE OF HFA [45] AND MSDA [46] MODELS ON MULTIPLE FEATURES DATASET OVER 10 SIMULATION RUNS.

Method	Source domain	Target domain fac	fou	kar	mor	pix	zer
HFA	**fac**	0.9707	0.5093	0.6737	0.623	0.7767	0.5900
	fou	0.7750	0.8043	0.6687	0.6137	0.7760	0.6000
	kar	0.7807	0.6000	0.9591	0.6144	0.7720	0.6600
	mor	0.7417	0.6500	0.5483	0.7384	0.6830	0.6900
	pix	0.7920	0.6300	0.6767	0.6193	0.9724	0.6600
	zel	0.7847	0.6500	0.6233	0.6107	0.7527	0.8199
mSDA	**fac**	0.9601	0.1840	0.2453	0.3000	0.6217	0.3743
	fou	0.6597	0.7953	0.3350	0.3003	0.6123	0.4147
	kar	0.6457	0.3170	0.9440	0.2810	0.5483	0.3487
	mor	0.5410	0.4083	0.2420	0.6976	0.5167	0.3230
	pix	0.6243	0.1733	0.3480	0.2313	0.9644	0.3823
	zel	0.6920	0.2697	0.2333	0.2943	0.5770	0.8031
RSP-KCCA	**fac**	0.9560	0.7683	0.8987	0.6510	0.9233	0.7190
	fou	0.8573	0.7993	0.8520	0.6360	0.7853	0.7860
	kar	0.9270	0.7780	0.9454	0.6583	0.9540	0.7947
	mor	0.7503	0.6857	0.8290	0.7344	0.7920	0.7347
	pix	0.9503	0.7817	0.9177	0.6253	0.9590	0.7773
	zer	0.8400	0.7837	0.8040	0.6340	0.8773	0.8063

In the context of semi-supervised learning the use of the manifold regularization has been well studied in [37]. It has also been shown to be effective for the domain adaptation problems in [6]. This regularization mechanism restricts the learned function to have similar values for similar instances. This is particularly useful for integrating the unlabeled instances into the learning framework. It should be mentioned that in general the question of existence of the manifold structure for a particular dataset is beyond the scope of this manuscript. However, an empirical indication of the existence and help of the integration of the manifold structure in the learning process can for instance be validated by the value of the tuned regularization parameter for the manifold regularization term in the cost function compared to the other regularization parameters. This analysis for the IXMAS dataset suggests that the manifold assumption might not hold completely for this dataset. Finally, the average test accuracy of different methods over Multiple features and IXMAS datasets are shown in Fig. 10. One may notice that for these datasets applying domain adaption models can considerably enhance the performance of the classifier.

IX. CONCLUSIONS

In this paper, we proposed a unified model using kernel canonical correlation analysis as core model for learning domain-invariant features used for cross-domain classification



(a) (b) (c)

Fig. 9. Analysis of the projected training data from source and target domains for the Multiple Features dataset. **(a)** The linear correlation between the projected training data from source and target domains using FS-KCCA model. **(b)** The linear correlation between the projected training data from source and target domains using SP-KCCA model with explicit feature map. **(c)** The correlation between the projected training data from source and target domains using RSP-KCCA model with explicit feature map.

TABLE III
COMPARING THE AVERAGE TEST ACCURACY OF THE PROPOSED RSP-KCCA APPROACH WITH THOSE OF HFA [45], GFK [3] AND MSDA [46] MODELS ON IXMAS ACTION DATASET OVER 10 SIMULATION RUNS.

Method	Source domain	Target domain				
		cam0	**cam1**	**cam2**	**cam3**	**cam4**
HFA	**cam0**	0.8035	0.6458	0.6591	0.6638	0.5606
	cam1	0.6657	0.7817	0.6572	0.6610	0.5634
	cam2	0.6714	0.6354	0.8134	0.6581	0.5606
	cam3	0.6629	0.6430	0.6544	0.8343	0.5559
	cam4	0.6610	0.6402	0.6496	0.6600	0.7775
GFK	**cam0**	0.6850	0.5896	0.6212	0.6326	0.5038
	cam1	0.6124	0.6787	0.6326	0.6465	0.5076
	cam2	0.6275	0.5922	0.6806	0.6490	0.5152
	cam3	0.6073	0.5846	0.6136	0.6995	0.5025
	cam4	0.6124	0.5960	0.6313	0.6338	0.6029
mSDA	**cam0**	0.7748	0.5096	0.5284	0.4688	0.4186
	cam1	0.5270	0.7568	0.5398	0.5133	0.4205
	cam2	0.5350	0.5104	0.8013	0.5000	0.3708
	cam3	0.5218	0.5019	0.5227	0.8201	0.4214
	cam4	0.5013	0.5242	0.4792	0.4494	0.7637
RSP-KCCA	**cam0**	0.7893	0.5966	0.6411	0.6657	0.5303
	cam1	0.6420	0.7718	0.6392	0.6127	0.5246
	cam2	0.6581	0.6496	0.8035	0.6449	0.5767
	cam3	0.6420	0.6458	0.6828	0.8423	0.5720
	cam4	0.6269	0.6430	0.6364	0.6506	0.7822

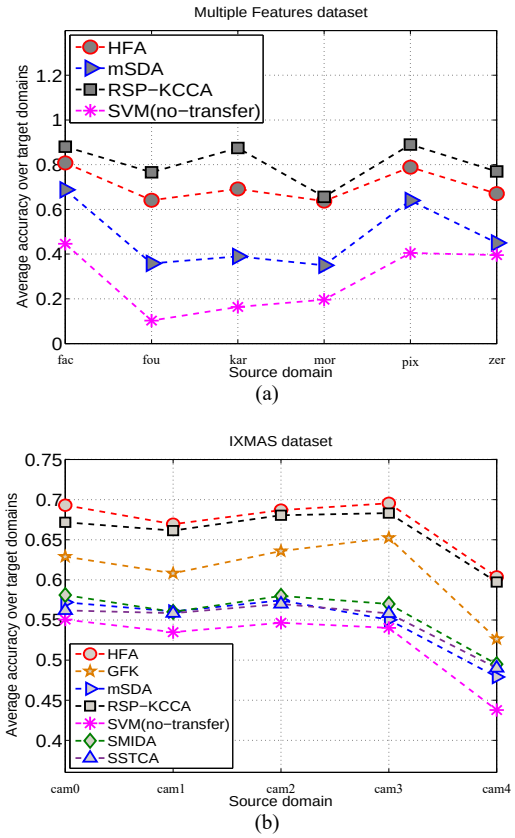


Fig. 10. Average test accuracy of different methods over Multiple features and IXMAS datasets. SVM method is applied on the original (not transferred) source data and its accuracy is tested on the original target data. For the SVM approach, the zero padding is applied to the input data whenever there is a dimensionality mismatch between the source and target domains. **(a)** Multiple feature dataset. **(b)** IXMAS dataset.

problem. The formulation is based on LS-SVMs and follows from a constrained optimization framework providing primal-dual insights. All possible side-information about the problem at hand, i.e. unlabeled, labeled, paired and unpaired instances are incorporated into the core model through proper regularization terms. Particularly interesting, we showed that adding labeled data instances into the formulation of KCCA in the primal, changes the eigenvalue problem associated with KCCA to a linear system of equations in the dual. In addition, the learned subspace preserves the label predictive information. In order to make the proposed approach scalable, we proposed to use Nyström approximation of the feature map and solve the optimization problem in the primal. The validity and applicability of the proposed methods are shown on synthetic and real-life datasets.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC AdG A-DATADRIVE-B (290923). This paper reflects only the authors' views, the Union is not liable for any use that may be made of the contained information; ● Research Council KUL: GOA/10/09 MaNet, CoE PFV/10/002 (OPTEC), BIL12/11T, several PhD/postdoc grants ● Flemish Government: ○ IOF: IOF/KP/SCORES4CHEM; ○ FWO: PhD/postdoc grants, projects: G.0377.12 (Structured systems), G.088114N (Tensor based data similarity); ○ IWT: PhD/Postdoc grants, projects: SBO POM (100031); ○ iMinds Medical Information Technologies SBO 2014; ● Belgian Federal Science Policy Office: IUAP P7/19 (DYSCO, Dynamical systems, control and optimization, 2012-2017).

Siamak Mehrkanoon is a postdoctoral research fellow at the KU Leuven, Belgium funded by postdoctoral mandates fellowship (PDM-3E150555: Advanced Data Driven Semi-Supervised Models and Applications). Johan Suykens is a full professor at the KU Leuven, Belgium.

REFERENCES

[1] S. J. Pan, J. T. Kwok, and Q. Yang, "Transfer learning via dimensionality reduction." in *AAAI*, vol. 8, 2008, pp. 677–682.

[2] J. Blitzer, R. McDonald, and F. Pereira, "Domain adaptation with structural correspondence learning," in *Proceedings of the 2006 conference on empirical methods in natural language processing*, 2006, pp. 120–128.

[3] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic f ow kernel for unsupervised domain adaptation," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2066–2073.

[4] H. Daumé III, "Frustratingly easy domain adaptation," *arXiv preprint arXiv:0907.1815*, 2009.

[5] L. Cheng and S. J. Pan, "Semi-supervised domain adaptation on manifolds," *Neural Networks and Learning Systems, IEEE Transactions on*, vol. 25, no. 12, pp. 2240–2249, 2014.

[6] J. Donahue, J. Hoffman, E. Rodner, K. Saenko, and T. Darrell, "Semi-supervised domain adaptation with instance constraints," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 668–675.

[7] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 199–210, 2011.

[8] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classif cation: A deep learning approach," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 513–520.

[9] Y.-R. Yeh, C.-H. Huang, and Y.-C. F. Wang, "Heterogeneous domain adaptation and classif cation by exploiting the correlation subspace," *Image Processing, IEEE Transactions on*, vol. 23, no. 5, pp. 2009–2018, 2014.

[10] M. Sugiyama, S. Nakajima, H. Kashima, P. V. Buenau, and M. Kawanabe, "Direct importance estimation with model selection and its application to covariate shift adaptation," in *Advances in neural information processing systems*, 2008, pp. 1433–1440.

[11] Y. Sun, M. S. Kamel, A. K. Wong, and Y. Wang, "Cost-sensitive boosting for classif cation of imbalanced data," *Pattern Recognition*, vol. 40, no. 12, pp. 3358–3378, 2007.

[12] G. Ditzler and R. Polikar, "An ensemble based incremental learning framework for concept drift and class imbalance," in *Neural Networks (IJCNN), The 2010 International Joint Conference on*. IEEE, 2010, pp. 1–8.

[13] J. Z. Kolter and M. Maloof, "Dynamic weighted majority: A new ensemble method for tracking concept drift," in *Third IEEE International Conference on Data Mining (ICML)*. IEEE, 2003, pp. 123–130.

[14] M. Sugiyama, M. Krauledat, and K.-R. Müller, "Covariate shift adaptation by importance weighted cross validation," *The Journal of Machine Learning Research*, vol. 8, pp. 985–1005, 2007.

[15] J. Jiang and C. Zhai, "Instance weighting for domain adaptation in nlp," in *ACL*, vol. 7, 2007, pp. 264–271.

[16] E. Tzeng, C. Devin, J. Hoffman, C. Finn, X. Peng, S. Levine, K. Saenko, and T. Darrell, "Towards adapting deep visuomotor representations from simulated to real environments," *arXiv preprint arXiv:1511.07111*, 2015.

[17] P. Prettenhofer and B. Stein, "Cross-language text classif cation using structural correspondence learning," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 1118–1127.

[18] Y. Zhu, Y. Chen, Z. Lu, S. J. Pan, G.-R. Xue, Y. Yu, and Q. Yang, "Heterogeneous transfer learning for image classif cation." in *AAAI*, 2011.

[19] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle, *Least squares support vector machines*. Singapore: World Scientif c Pub. Co., 2002.

[20] P. L. Lai and C. Fyfe, "Kernel and nonlinear canonical correlation analysis," *International Journal of Neural Systems*, vol. 10, no. 05, pp. 365–377, 2000.

[21] F. R. Bach and M. I. Jordan, "Kernel independent component analysis," *The Journal of Machine Learning Research*, vol. 3, pp. 1–48, 2003.

[22] G. H. Golub and C. F. Van Loan, *Matrix computations*. JHU Press, 2012, vol. 3.

[23] C. Williams and M. Seeger, "Using the Nyström method to speed up kernel machines," in *Advances in Neural Information Processing Systems 13*, 2001.

[24] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Advances in neural information processing systems*, 2007, pp. 1177–1184.

[25] T. Yang, Y.-F. Li, M. Mahdavi, R. Jin, and Z.-H. Zhou, "Nyström method vs random fourier features: A theoretical and empirical comparison," in *Advances in neural information processing systems*, 2012, pp. 476–484.

[26] Y.-J. Lee and O. L. Mangasarian, "Rsvm: Reduced support vector machines." in *SDM*, vol. 1, 2001, pp. 325–361.

[27] S. Mehrkanoon and J. A. K. Suykens, "Large scale semi-supervised learning using KSC based model," in *In proc. of the International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2014, pp. 4152–4159.

[28] C. T. Baker and C. Baker, *The numerical treatment of integral equations*. Clarendon press Oxford, 1977, vol. 13.

[29] M. Girolami, "Orthogonal series density estimation and the kernel eigenvalue problem," *Neural Computation*, vol. 14, no. 3, pp. 669–688, 2002.

[30] A. Kimura, H. Kameoka, M. Sugiyama, T. Nakano, E. Maeda, H. Sakano, and K. Ishiguro, "Semicca: Eff cient semi-supervised learning of canonical correlations," in *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 2010, pp. 2933–2936.

[31] M. B. Blaschko, C. H. Lampert, and A. Gretton, "Semi-supervised Laplacian regularization of kernel canonical correlation analysis," in *Machine Learning and Knowledge Discovery in Databases*. Springer, 2008, pp. 133–145.

[32] M. Volpi, G. Matasci, M. Kanevski, and D. Tuia, "Semi-supervised multiview embedding for hyperspectral data classif cation," *Neurocomputing*, vol. 145, pp. 427–437, 2014.

[33] S. Mehrkanoon, O. M. Agudelo, and J. A. K. Suykens, "Incremental multi-class semi-supervised clustering regularized by kalman f ltering," *Neural Networks*, vol. 71, pp. 88–104, 2015.

[34] S. Mehrkanoon, C. Alzate, R. Mall, R. Langone, and J. A. K. Suykens, "Multiclass semisupervised learning based upon kernel spectral clustering," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 4, pp. 720–733, 2015.

[35] S. Mehrkanoon, T. Falck, and J. A. K. Suykens, "Approximate solutions to ordinary differential equations using least squares support vector machines," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 9, pp. 1356–1367, 2012.

[36] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

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

[38] O. Chapelle, B. Schölkopf, A. Zien *et al.*, "Semi-supervised learning," 2006.

[39] T. Sun, S. Chen, J. Yang, and P. Shi, "A novel method of combined feature extraction for recognition," in *Eighth IEEE International Conference on Data Mining*. IEEE, 2008, pp. 1043–1048.

[40] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 251–260.

[41] A. Asuncion and D. J. Newman, "UCI machine learning repository," 2007.

[42] B. Kulis, K. Saenko, and T. Darrell, "What you saw is not what you get: Domain adaptation using asymmetric kernel transforms," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1785–1792.

[43] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer vision and image understanding*, vol. 104, no. 2, pp. 249–257, 2006.

[44] R. Li and T. Zickler, "Discriminative virtual views for cross-view action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 2855–2862.

[45] W. Li, L. Duan, D. Xu, and I. W. Tsang, "Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 6, pp. 1134–1148, 2014.

[46] M. Chen, Z. Xu, K. Weinberger, and F. Sha, "Marginalized denoising autoencoders for domain adaptation," in *Proceedings of the 29th International Conference on International Conference on Machine Learning*. ICML, 2012, pp. 1627–1634.

[47] K. Yan, L. Kou, and D. Zhang, "Learning domain-invariant subspace using domain features and independence maximization," *IEEE transactions on cybernetics*, 2017.