Investigating Feature Extraction for Domain Adaptation in Remote Sensing Image Classification

Giona Matasci[1], Lorenzo Bruzzone[2], Michele Volpi[1], Devis Tuia[3] and Mikhail Kanevski[1]
[1]*Center for Research on Terrestrial Environment, University of Lausanne, Lausanne, Switzerland*
[2]*Remote Sensing Laboratory, University of Trento, Trento, Italy*
[3]*Laboratory of Geographic Information Systems, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland*

Keywords: Satellite Imagery, Image Classification, Transfer Learning, Manifold Alignment, Kernel Methods.

Abstract: In this contribution, we explore the feature extraction framework to ease the knowledge transfer in the thematic classification of multiple remotely sensed images. By projecting the images in a common feature space, the purpose is to statistically align a given target image to another source image of the same type for which we dispose of already collected ground truth. Therefore, a classifier trained on the source image can directly be applied on the target image. We analyze and compare the performance of classic feature extraction techniques and that of a dedicated method issued from the field of domain adaptation. We also test the influence of different setups of the problem, namely the application of histogram matching and the origin of the samples used to compute the projections. Experiments on multi- and hyper-spectral images reveal the benefits of the feature extraction step and highlight insightful properties of the different adopted strategies.

1 INTRODUCTION

In the field of remote sensing, when dealing with the supervised thematic classification of a given image, the availability of labeled samples from other acquisitions can alleviate the effort associated with the ground truth collection task. Therefore, procedures allowing a classifier trained on one image, the *source image*, to perform efficiently on a different but related image (same sensor and set of classes), the *target image*, are highly demanded by the users [Bruzzone and Prieto, 2001]. These techniques could limit the expensive field campaigns or time-consuming photo interpretation analyses needed to define a training set when previously obtained information linking spectral signatures and ground cover classes is not available. However, there can be heavy radiometric differences between the images due to varying illumination and atmospheric conditions, seasonal effects affecting the vegetation, changing acquisition geometry, etc. These factors induce a shift in the statistical distribution of the land cover spectra.

To address this issue and make the images more similar to each other, the basic approaches involve the use of demanding physical models (*e.g.* atmospheric compensations) or very simple signature extension approaches [Woodcock et al., 2001]. Recently, other more sophisticated strategies, relying on the statistical properties of the analyzed datasets, have been proposed. To improve the standard univariate PDF matching procedure of *histogram matching* (HM), in [Inamdar et al., 2008] the authors propose its multivariate extension. Such a procedure is designed to take into account the correlation between bands. In [Tuia et al., 2012], a correspondence between the data manifolds is sought by means of graphs in order to deform and align the images.

In the pattern recognition and machine learning communities, the above-mentioned problems are studied in the framework known as *domain adaptation* (DA) [Pan and Yang, 2010]. Among the DA methods, we find a set of techniques aimed at transferring the knowledge via the so-called *feature-representation-transfer* approach. The goal of this type of procedures is to build a set of shared and invariant features, either by *feature extraction* (FE) or by *feature selection*, which are able to reduce the differences of statistical distribution between the two domains.

Subsequently, one is enabled to apply a model trained on the source image to classify another target image of interest. The same line of reasoning applies to localized reference data, only partially covering the complete class distribution. When these data have to

Matasci G., Bruzzone L., Volpi M., Tuia D. and Kanevski M. (2013).
Investigating Feature Extraction for Domain Adaptation in Remote Sensing Image Classification.
In *Proceedings of the 2nd International Conference on Pattern Recognition Applications and Methods*, pages 419-424
DOI: 10.5220/0004199504190424

be used to generalize over the entire image, a *sample selection bias* problem is likely to occur. In remote sensing, the two aforementioned DA problems are addressed by *partially unsupervised* or *semi-supervised* classification tasks.

In the literature, while a wealth of different FE methods have been applied to single images [Arenas-García and Petersen, 2009], few papers tackle the simultaneous analysis of multiple remotely sensed images through dimensionality reduction. In [Nielsen et al., 1998], the authors introduce a method, based on canonical correlation analysis, to detect changes in bi-temporal images. The technique aims at projecting the samples into a space where the extracted components display similar values for the unchanged regions while maximally differing on the changed ones. However, this methodology is restricted to the study of spatially co-registered images. The selection of invariant features is investigated in [Bruzzone and Persello, 2009]. It has been proven that, when working with a hyperspectral image, it is possible to select a discriminant subset of the numerous bands that bears the highest spatial invariance across the image to improve the generalization abilities of a classifier.

In the present contribution we study the application of FE techniques to reduce the distribution divergence between source and target domains while keeping the main data properties. We study their impact when implemented in a cross-domain setting in combination with the widely used HM procedure. Starting with images having either unmatched (original) or matched histograms, prior to the classification task, FE is performed on a subset of pixels coming either from a single or from both images. Once the projection is defined, a common identical mapping of the images is carried out. In the new feature space, we should observe: 1) datasets displaying more similar probability distributions and 2) more separable thematic classes. Then, a simple supervised classifier learned on the source image, where the pixels have been sampled, could be used to predict the target image. In our experiments, keeping fixed the base classifier, we compare the effectiveness of FE via *Principal Component Analysis* (PCA), *Kernel Principal Component Analysis* (KPCA) and *Transfer Component Analysis* (TCA), which is a procedure especially designed for DA. Additionally, we investigate the influence of other factors affecting the knowledge transfer process, such as the origin (source image only or both images) of the pixels used to define the projection or the nature (linear or non-linear) of the classification model.

2 DOMAIN ADAPTATION VIA FEATURE EXTRACTION

Let $\mathcal{D}_S = \{X_S, Y_S\} = \{(\mathbf{x}_{S_i}, y_{S_i})\}_{i=1}^{n_s}$ be the set of n_s labeled source training data and $X_T = \{\mathbf{x}_{T_j}\}_{j=1}^{n_t}$ the set of the n_t unlabeled target data, with samples $\mathbf{x}_{S_i}, \mathbf{x}_{T_j} \in \mathbb{R}^d \, \forall i, j$. The goal of the partially unsupervised approaches considered in this paper is to predict labels $y_{T_j} \in \Omega = \{\omega_c\}_{c=1}^{C}$ (set of C classes in common with $\mathcal{D}_S$) based exclusively on the use of labeled data from $\mathcal{D}_S$ in the training phase. To this end, a common mapping ϕ of the samples of both domains is needed such that $P(X_S^*) \approx P(X_T^*)$, with $X_S^* = \phi(X_S), X_T^* = \phi(X_T)$. In practice, we need a matrix $\mathbf{W}$ to perform the joint mapping ϕ of the data. This mapping matrix can be found based on a subset of samples X from either

- the two domains, *i.e.* $X \subseteq X_S \cup X_T$, or
- one domain only, *i.e.* $X \subseteq X_S$ (or X_T).

Standard FE methods can be employed to estimate $\mathbf{W}$ and embed data in a m-dimensional space with $m \ll d$. In the next sections, we will briefly illustrate two techniques for non-linear FE.

2.1 Kernel Principal Component Analysis

Kernel PCA [Schölkopf et al., 1998], the non-linear counterpart of standard PCA, aims at extracting a set of features or components onto which it projects the original data to improve their representation.

Let us consider the $n \times d$ matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top$ composed of the n column vectors $\mathbf{x}_i \in \mathbb{R}^d$ belonging to dataset X (centered to zero mean). Classical PCA aims at finding the directions of maximal variance (*i.e.* diagonalizing the covariance matrix) by solving the following eigenproblem (primal formulation)

$$\frac{1}{n-1}\mathbf{X}^\top \mathbf{X}\mathbf{u} = \lambda \mathbf{u}\,. \tag{1}$$

It is possible to show that the corresponding dual formulation leading to KPCA

$$\frac{1}{n-1}\mathbf{X}\mathbf{X}^\top \alpha = \lambda \alpha \tag{2}$$

yields the same non-zero eigenvalues λ and that its eigenvectors α are related to their primal counterparts $\mathbf{u}$.

By applying the well-known *kernel trick* in order to implicitly simulate a mapping φ of the samples into a higher-dimensional Reproducing Kernel Hilbert Space (RKHS), Eq. (2) becomes

$$\begin{array}{rcl} \frac{1}{n-1}\varphi(\mathbf{X})\varphi(\mathbf{X})^\top \alpha & = & \lambda\alpha \quad \Leftrightarrow \\ \frac{1}{n-1}\mathbf{K}\alpha & = & \lambda\alpha\,, \end{array} \tag{3}$$

where $\mathbf{K}$ is the kernel matrix of elements $K_{i,j} = \varphi(\mathbf{x}_i)^\top \varphi(\mathbf{x}_j)$. Dropping the $1/(n-1)$ factor and by using the centered kernel matrix $\tilde{\mathbf{K}} = \mathbf{HKH}$, with centering matrix $\mathbf{H} = \mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n^\top/n$, the final KPCA eigenvalue problem is set up as

$$\tilde{\mathbf{K}}\alpha = \lambda\alpha. \tag{4}$$

The resulting projection of some test samples $\mathbf{X}_{\text{test}}$ (*e.g.* the complete images) on the first *m kernel principal components* is expressed as $\mathbf{X}^*_{\text{test}} = \tilde{\mathbf{K}}_{\text{test}}\mathbf{W}$, where $\tilde{\mathbf{K}}_{\text{test}}$ is the centered test kernel and $\mathbf{W}$ is constituted by the first m eigenvectors $[\alpha_1, \ldots, \alpha_m]$.

2.2 Transfer Component Analysis

The other kernel-based FE technique we tested is especially designed for DA. In fact, the TCA method [Pan et al., 2011] aims at finding a common embedding of the data from the two domains that minimizes the divergence between the distributions. To estimate this shift, TCA resorts to a recently proposed measure, the *Maximum Mean Discrepancy* (MMD) [Borgwardt et al., 2006]. This is a non-parametric, kernel-based, multivariate measure of divergence between probability distributions.

The empirical estimate of the MMD between distributions of a given source dataset X_S and target dataset X_T is computed as

$$\text{MMD}(X_S, X_T) = \text{Tr}(\mathbf{KL}), \tag{5}$$

where

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{S,S} & \mathbf{K}_{S,T} \\ \mathbf{K}_{T,S} & \mathbf{K}_{T,T} \end{bmatrix} \in \mathbb{R}^{(n_s+n_t)\times(n_s+n_t)}, \tag{6}$$

with $\mathbf{K}_{S,S}, \mathbf{K}_{T,T}, \mathbf{K}_{S,T}, \mathbf{K}_{T,S}$ being the kernel matrices obtained from the data of the source domain, target domain and cross domains, respectively. Moreover, if $\mathbf{x}_i, \mathbf{x}_j \in X_S$, then $L_{i,j} = 1/n_s^2$, else if $\mathbf{x}_i, \mathbf{x}_j \in X_T$ we have $L_{i,j} = 1/n_t^2$, otherwise, $L_{i,j} = -1/n_s n_t$. We interpret MMD as the squared distance between the means, computed in the feature space, of the samples belonging to the two domains. This quantity equals zero when the two distributions are exactly the same.

The purpose of the TCA algorithm is to find a mapping function ϕ, and thus a projection matrix $\mathbf{W} \in \mathbb{R}^{(n_s+n_t)\times m}$ (with $m \ll n_s + n_t$), that is able to reduce the distance between the probability distributions of $\phi(X_S)$ and $\phi(X_T)$ (MMD minimization) while preserving the main properties of the original data X_S and X_T (maximization of data variance as in PCA and KPCA).

The kernel learning problem solved by TCA is

$$\begin{aligned} \min_{\mathbf{W}} \quad & \left\{ \text{Tr}(\mathbf{W}^\top \mathbf{KLKW}) + \mu \text{Tr}(\mathbf{W}^\top \mathbf{W}) \right\} \\ \text{s.t.} \quad & \blacksquare^* = \mathbf{I}_m . \end{aligned} \tag{7}$$

The first term is the MMD between mapped samples $\text{MMD}(X^*_S, X^*_T)$, which should thus be minimized according to the TCA objectives. The second one is a regularizer controlling the complexity of $\mathbf{W}$, whose influence is tuned by the tradeoff parameter μ. The constraint is used to enforce variance maximization, which is the other goal of TCA. Indeed, $\blacksquare^* = \mathbf{W}^\top \tilde{\mathbf{K}} \mathbf{W}$ is the covariance matrix of the data in the projection space which is constrained to orthogonality by the identity matrix $\mathbf{I}_m$.

The problem in (7) can be reformulated as a trace maximization problem whose solution yields the mapping matrix $\mathbf{W}$ through the eigendecomposition of

$$\mathbf{M} = (\mathbf{KLK} + \mu\mathbf{I})^{-1}\mathbf{KHK}, \tag{8}$$

and keeping the m eigenvectors associated with the m largest eigenvalues $\text{eig}(\mathbf{M})$.

Finally, we compute the *m transfer components* for new test samples $\mathbf{X}_{\text{test}}$ as $\mathbf{X}^*_{\text{test}} = \mathbf{K}_{\text{test}}\mathbf{W}$, where $\mathbf{K}_{\text{test}}$ is the test kernel.

3 DATA DESCRIPTION AND EXPERIMENTAL SETUP

3.1 Datasets

The first dataset used for the experiments is the 1.3 m spatial resolution image acquired by the ROSIS-03 hyperspectral sensor over the city of Pavia, Italy. The 102 retained bands cover a region of the spectrum between 0.43 and 0.86 μm. In this urban setting, 4 classes have been taken into account: "buildings", "roads", "shadows" and "vegetation". Because of different materials constituting the roofs as well as the roads and due to the different types of vegetation, the spectral signatures of these ground cover classes bear a remarkable variation across the image.

Thus, we considered two spatially disjoint subsets of the scene to assess the ability of the different FE techniques in transferring the knowledge: a source sub-region of 172×123 pixels and a target sub-region 350×350 pixels. The spatial extent of the starting source sub-image is quite small, raising the question of the representativity of the training samples while generalizing over the Pavia scene (simulated sample selection bias problem). Indeed, the description of the classes is presumably not rich enough to account for the complete variation of the spectral signatures. The dataset shift level is here deemed to be light.

The second dataset consists of two VHR QuickBird images of two different neighborhoods of the city of Zurich, Switzerland, acquired in August 2002

and in October 2006. For the empirical assessment of the techniques, we defined the image of 2006 as being the source image while taking the 2002 image as the target image. The shift occurred between the two acquisitions is judged as large in this case. In fact, we notice differences in illumination conditions due to the sun elevation and acquisition geometry, seasonal effects affecting the vegetation and a different nature of the materials used for roofs and roads. The standard 4 QuickBird bands in the VNIR spectrum (450 to 900 nm) have been completed by textural and morphological features to reach a final set of 16 features. For the classification task we defined 5 classes found on both images: "buildings", "roads", "grass", "trees" and "shadows".

For the two datasets, the variables have been normalized to zero mean and unit variance, based on the source image descriptive statistics.

3.2 Design of the Experiments

In order to comprehensively assess the advantages of the different FE methods when combined with linear or non-linear models, we chose *Linear Discriminant Analysis* (LDA) and *Quadratic Discriminant Analysis* (QDA) as base classifiers.

For the key FE step we applied the 3 mentioned techniques: PCA, KPCA and TCA. The σ parameter of the Gaussian RBF kernel, used for both the KPCA and TCA, has been set as the median distance among the data points. A sensitivity analysis and other previous works [Pan et al., 2011], suggested to set to 1 the value for the TCA tradeoff parameter μ. The classification models have been trained with source samples mapped into a space of increasing dimension (1 to 18 or 15 features for the Pavia and Zurich datasets, respectively). PCA and KPCA have been run in 3 different settings. First, the mapping matrix $\mathbf{W}$ has been computed based on samples coming from both images (standard setting). A second test involved a FE on the source image alone, with a subsequent identical mapping of the target image (same $\mathbf{W}$ used for the projection of both domains). The third approach considered a separate, independent, mapping of the two domains (different $\mathbf{W}$). In this setting, just the results with PCA are reported. TCA was only run in the first setting, since this technique is explicitly designed to handle data issued from two different domains. As upper and lower bounds, classifiers trained with samples only belonging to the target or source image have also been tested. In these cases, the input space was constituted by the original spectral bands (plus spatial information for the Zurich images). A summary of all these settings with related names is reported in Tab. 1.

Table 1: Methods and settings compared in the experiments using either LDA (`L`) or QDA (`Q`) as classifiers.

Name	FE method	FE based on	Classifier trained on
(L/Q)DAtgt	-	-	target im.
(L/Q)DAsrc	-	-	source im.
(L/Q)DA_PCA	PCA	both im.	source im.
(L/Q)DA_PCA_1DOM	PCA	source im.	source im.
(L/Q)DA_PCA_INDEP	PCA	both im. indep.	source im.
(L/Q)DA_KPCA	KPCA	both im.	source im.
(L/Q)DA_KPCA_1DOM	KPCA	source im.	source im.
(L/Q)DA_TCA	TCA	both im.	source im.

The influence of the HM procedure as a preprocessing step has also been investigated. The series of experiments depicted above has been carried out without and with the univariate match of the distributions of the two images (source image as reference). To capture the hypothetic loss in accuracy when repredicting on the source image after having extracted the features using data from both images, classification performances on the source images have also been recorded.

For both datasets, 200 pixels per class have been retained to build the training sets. The set of unlabeled target pixels used to compute the projection counted $200 \cdot C$ pixels randomly selected all over the corresponding image. Experiments with 10 independent realizations of these sets have been run to ensure a fair comparison.

4 RESULTS AND DISCUSSION

4.1 Pavia ROSIS Dataset

The left panel of Fig. 1 depicts the performance of the LDA on the Pavia target image. Fig. 1(a) reports the results obtained on the raw images, whereas Fig. 1(b) shows the behavior after HM. One can notice the large gap between in-domain (`LDAtgt`: solid blue line) and out-domain (`LDAsrc`: dashed red line) models existing in both plots. Nonetheless, the impact of the HM as a preprocessing step is quite remarkable. In fact, LDA models trained on original target data outperform LDA models based on original source data by 0.356 κ points when no matching is performed, while this difference reduces to 0.188 κ points after matching.

In between these reference lines, we observe two distinct trends. The first one concerns kernel-based FE methods (`LDA_KPCA`: dashed purple line, `LDA_KPCA_1DOM`: solid light green line, `LDA_TCA`: dashed black line) that yield a robust performance with accuracies reaching and even exceeding those of

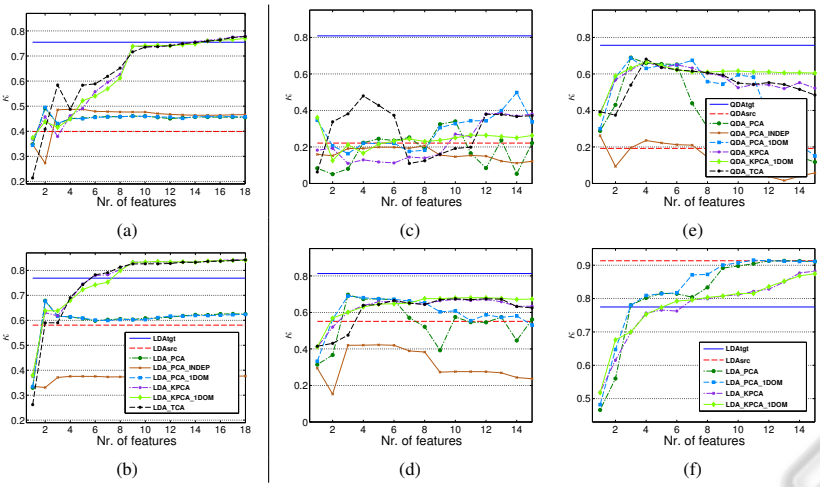

Figure 1: Classification performances (average of estimated κ statistic over 10 runs) on the (left) Pavia and (right) Zurich datasets considering several different settings. Target domain test sets included 14'047 (Pavia) and 26'797 (Zurich) samples. (a) LDA on the Pavia target image, without HM. (b) LDA on the Pavia target image, with HM. (c) LDA on the Zurich target image, without HM. (d) LDA on the Zurich target image, with HM. (e) QDA on the Zurich target image, with HM. (f) LDA on the Zurich source image after FE, with HM (test set of 12'310 pixels). Legend of (b) also valid for (a), (c) and (d).

the target models when using at least 14 (no HM) or 8 (with HM) extracted features. After these thresholds, the 3 techniques converge to very similar performances, indicating the non-inferiority of KPCA with respect to a domain adaptation technique as TCA. Such a behavior also suggests that basing the FE on one domain only (the source image) does not imply a loss in invariance across domains. Indeed, rather than the reduction of the statistical divergence between datasets (as measured by the MMD), it seems that the extraction of features provides larger benefits in terms of class discrimination. The latter is highly increased in the two domains, especially when resorting to kernel-based methods, easing thus the drawing of meaningful and domain invariant class boundaries. Note that the feature extractors employed in our tests do not explicitly aim at optimizing class separation: this may be interpreted as an implicit benefit of the non-linear mapping.

The second trend is related to PCA-based methods (`LDA_PCA`: dashed dark green line, `LDA_PCA_1DOM`: dashed light blue line), which reveal a less satisfactory performance, just above the baseline of the `LDAsrc` model. Peak accuracies are obtained in both experiments with 2 features, while after, as noisy components come into play, the quality of the LDA model decreases. Also in this case, no difference is noticeable between the use of both domains for FE versus the use of the source domain only.

4.2 Zurich QuickBird Dataset

When considering the second dataset, Figs. 1(c) -(d) confirm the usefulness of HM. All the methods/settings tested failed if applied to unmatched data. Another key finding is the complementarity of the two pre-classification procedures. On both datasets, we noticed that the best accuracies are those reached by models built on images with matched histograms having undergone the FE. After these steps, the images are sufficiently aligned and the features are discriminant enough to allow classifiers trained on the source image to generalize well on the target image too.

Looking in details at Fig. 1(d), we witness a similar behavior as with the Pavia dataset. Kernel-based techniques need more features to attain good performances with respect to PCA. On this dataset, nevertheless, the best classification accuracy reached by both families of methods is comparable and still 0.1 κ points below the reference of the target domain model. Additionally, let us remark the slight superiority of the setting in which the FE is done exclusively on the source image (`LDA_PCA_1DOM`, `LDA_KPCA_1DOM`) with respect to an extraction based on both domains (`LDA_PCA`, `LDA_KPCA`). This trend, which is observable also on the previously examined dataset, was not expected, revealing some interesting properties of the tested approaches. Finally, as for the Pavia image, we observe the complete, though expected, failure of the `LDA_PCA_INDEP` approach (solid brown line), with an accuracy curve evolving far below the rest of the curves throughout the entire feature set.

Fig. 1(e) describes the behavior of the same alignment strategies, after HM, but when a non-linear classifier is used. The QDA curves depicted here show that the tendencies highlighted for linear models are

valid in this situation as well. It is worth noting the remarkable discriminant and invariant properties of the all the features extracted by KPCA from the source image. The QDA_KPCA_1DOM curve is the most stable across the entire range of features provided to the model.

In conclusion, Fig. 1(f) uncovers the behavior of some of the LDA models when asked, after HM and after the projection, to predict the class labels back on the source image. Although the pattern is not as evident as expected, we can appreciate the loss in accuracy induced by the FE based also on pixels issued from another domain. This confirms that out-domain data interfere with the proper extraction of discriminant domain-specific features, while improving the overall generalization abilities of the system when dealing with cross-domain knowledge transfer.

5 CONCLUSIONS

In this paper, the analysis of feature extraction techniques to jointly transform two related remote sensing images to align their feature spaces has been presented. After the projection, the matched images display an increased discrimination between ground cover classes, allowing a supervised classifier to obtain an accurate generalization on both source and target domains.

Experiments proved that the combination of the histogram matching procedure with the feature extraction step is extremely beneficial, confirming the mandatory application of the former before any domain adaptation task. Among the extraction techniques, we noticed the slight superiority of kernel-based features extractors (KPCA and TCA) with respect to simple linear techniques such as PCA. No notable differences have been observed between the two kernel methods. This fact suggests that, rather than the reduction of the divergence between marginal distributions governing the two images, as pursued by TCA, the key benefit is the increased class separability. Also, we found that the use of pixels from one image only to compute the projection provides equally invariant features as a joint sampling of the images.

These results open a number of opportunities to practitioners of the field dealing with large scale land cover mapping applications involving several remotely sensed images.

As an outlook on new research directions, we plan to test supervised FE methods. Techniques such as Kernel Fisher Discriminant Analysis, Kernel Canonical Correlation Analysis, Kernel Orthogonal Partial Least Squares, etc. could be used to find the proper projections based on the labeled source domain data.

ACKNOWLEDGEMENTS

This work has been supported by the Swiss National Science Foundation with grants no. 200021-126505 and PZ00P2-136827.

REFERENCES

Arenas-García, J. and Petersen, K. B. (2009). Kernel multivariate analysis in remote sensing feature extraction. In Camps-Valls, G. and Bruzzone, L., editors, *Kernel Methods for Remote Sensing Data Analysis*. J. Wiley & Sons, NJ, USA.

Borgwardt, K. M., Gretton, A., Rasch, M. J., Kriegel, H.-P., Schölkopf, B., and Smola, A. J. (2006). Integrating structured biological data by Kernel Maximum Mean Discrepancy. *Bioinformatics*, 22(14):e49–e57.

Bruzzone, L. and Persello, C. (2009). A novel approach to the selection of spatially invariant features for the classification of hyperspectral images with improved generalization capability. *IEEE Trans. Geosci. Remote Sens.*, 47(9):3180–3191.

Bruzzone, L. and Prieto, D. F. (2001). Unsupervised retraining of a maximum likelihood classifier for the analysis of multitemporal remote sensing images. *IEEE Trans. Geosci. Remote Sens.*, 39(2):456–460.

Inamdar, S., Bovolo, F., and Bruzzone, L. (2008). Multidimensional probability density function matching for preprocessing of multitemporal remote sensing images. *IEEE Trans. Geosci. Remote Sens.*, 46(4):1243–1252.

Nielsen, A. A., Conradsen, K., and Simpson, J. J. (1998). Multivariate alteration detection (MAD) and MAF postprocessing in multispectral, bitemporal image data: New approaches to change detection studies. *Remote Sens. Environ.*, 64:1–19.

Pan, S. J., Tsang, I., Kwok, J. T., and Yang, Q. (2011). Domain adaptation via transfer component analysis. *IEEE Trans. Neural Netw.*, 22(2):199–210.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(10):1345–1359.

Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 10(5):1299–1319.

Tuia, D., Muñoz-Marí, J., Gomez-Chova, L., and Malo, J. (2012). Graph matching for adaptation in remote sensing. *IEEE Trans. Geosci. Remote Sens.*, PP(99):1–13.

Woodcock, C. E., Macomber, S. A., Pax-Lenney, M., and Cohen, W. B. (2001). Monitoring large areas for forest change using Landsat: Generalization across space, time and Landsat sensors. *Remote Sens. Environ.*, 78(1-2):194–203.