remote sensing

MDPI

Article

Geodesic Flow Kernel Support Vector Machine for Hyperspectral Image Classification by Unsupervised Subspace Feature Transfer

Alim Samat [1,2,*], Paolo Gamba [3], Jilili Abuduwaili [1,2], Sicong Liu [4] and Zelang Miao [5]

[1] State Key Laboratory of Desert and Oasis Ecology, Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Urumqi 830011, China; jilil@ms.xjb.ac.cn

[2] Chinese Academy of Sciences Research Center for Ecology and Environment of Central Asia, Urumqi 830011, China

[3] Department of Electrical, Computer and Biomedical Engineering, University of Pavia, 27100 Pavia, Italy; paolo.gamba@unipv.it

[4] College of Surveying and Geoinformatics, Tongji University, Shanghai 200092, China; sicongliu.rs@gmail.com

[5] Department of Land Surveying and Geo-Informatics, Hong Kong Polytechnic University, Kowloon, Hong Kong 999077, China; cumtzlmiao@gmail.com

* Correspondence: alim.smt@gmail.com; Tel.: +86-991-788-5432

Academic Editors: Lenio Soares Galvao, Ioannis Gitas and Prasad S. Thenkabail
Received: 15 January 2016; Accepted: 4 March 2016; Published: 16 March 2016

Abstract: In order to deal with scenarios where the training data, used to deduce a model, and the validation data have different statistical distributions, we study the problem of transformed subspace feature transfer for domain adaptation (DA) in the context of hyperspectral image classification via a geodesic Gaussian flow kernel based support vector machine (GFKSVM). To show the superior performance of the proposed approach, conventional support vector machines (SVMs) and state-of-the-art DA algorithms, including information-theoretical learning of discriminative cluster for domain adaptation (ITLDC), joint distribution adaptation (JDA), and joint transfer matching (JTM), are also considered. Additionally, unsupervised linear and nonlinear subspace feature transfer techniques including principal component analysis (PCA), randomized nonlinear principal component analysis (rPCA), factor analysis (FA) and non-negative matrix factorization (NNMF) are investigated and compared. Experiments on two real hyperspectral images show the cross-image classification performances of the GFKSVM, confirming its effectiveness and suitability when applied to hyperspectral images.

Keywords: transfer learning; domain adaptation; geodesic flow kernel support vector machine; randomized nonlinear principal component analysis; feature transfer; image classification

1. Introduction

In the past few years, pattern recognition (PR) and machine learning (ML) techniques have been employed to derive land use/land cover (LULC) maps using remotely sensed (RS) data in automatic or semi-automatic ways [1–3]. Supervised classification methods ideally require a large amount of labeled samples with good statistical properties, as they need to be class unbiased and allow a good discrimination among classes. However, such labeled sample collection is a challenging task, especially for large area mapping and classification of multi-temporal high-resolution and hyperspectral remote sensing images [4,5]. Aiming at a more accurate classification with small training sets, many supervised [6], hybrid [7] and semi-supervised learning (SSL) [8,9] methods have been proposed for multi/hyper-spectral image classification in recent years.

Furthermore, in real applications one is often faced with scenarios where the training data used to deduce a model and the validation data have different statistical distributions. As a matter of fact, radiometric differences, atmospheric and illumination conditions, seasonal variation, and variable acquisition geometries may cause such distribution shifts in single or multi-temporal RS images [5,7,10]. Existing inconsistencies between the source and target data distributions include covariate shift [11] or sampling selection bias [12]. To tackle such distribution shifts, it is very difficult and expensive to recollect enough training data. Traditional solutions, including image-to-image normalization [13], absolute and relative image normalization [14,15], histogram matching (HM) [16], and a multivariate extension of the univariate matching [17], have been therefore considered. However, a more efficient solution would be to transfer knowledge between the domains. In this regard, transfer learning (TL) and domain adaptation (DA) methods have recently attracted ever increasing attention [18–21].

According to the technical literature, transfer learning can be categorized as inductive transfer learning (ITL), transductive transfer learning (TTL) and unsupervised transfer learning (UTL) [22] according to the availability of labeled data only for the target domain, or for the source domain, or their total unavailability. Multi-task learning (MTL) is a special case of ITL, where labeled data is available for the source and target domains, and the system is trained simultaneously in both of them [23]. Similarly, domain adaptation is a special case of TTL, when labeled data are available only in the source domain, but there is only a single task to be learned [22,24]. Finally, based on the kind of "knowledge" transferred across domains, transfer learning can be roughly clustered into instance-based transfer learning (e.g., instance reweighting [25,26] and importance sampling [27]), feature-based transfer learning [28], parameter transfer approaches [29], and relational knowledge transfer techniques [30].

In this work, we focus on domain adaptation (DA). Seminal works, such as the adjustment of parameters for a maximum-likelihood classifier in a multiple cascade classifier system by retraining, have been carried out in [5,7] to update a land-cover map using multi-temporal RS images. Multivariate alteration detection (MAD) was introduced in [31], and was proved to be effective in detecting invariant regions in bi-temporal images. Other DA methods, such as the feature selection approach [32], manifold transformation [33], multi-domain method [34] and domain adaptation support vector machine (DASVM) [18] have also been proposed. A semi-supervised transfer component analysis (SSTCA) method for DA [35] was thoroughly investigated in [21] for hyperspectral and high resolution image classification. Finally, considering that labeled samples may be insufficiently available in some cases, active learning (AL) has been also combined with DA. For instance, an SVM-based boosting of AL strategies for efficient DA in very high resolution (VHR) and hyperspectral image classification was proposed in [36].

As mentioned above, one possible way to transfer knowledge between two domains is to look for a feature-based representation of the source and target domains that minimizes cross domain differences [22,35,37]. To this aim, feature extraction (FE) methods are used to project the two domains into a common latent space, implementing transfer component analysis (TCA) [21,35]. For instance, sampling geodesic flow (SGF) and geodesic flow kernel (GFK) were proposed in [38,39] to exploit subspaces in a Grassmannian manifold [40].

Notwithstanding the many studies in this area, challenging issues still remain open. For example, due to the huge computation complexity of the empirical estimate of maximum mean discrepancy (MMD) between the distribution of source and target domains, TCA is only suitable for small images. Similarly, the SGF approach requires to select *a priori* a sampling strategy, the transformed features to consider, and the dimensionality of the transformed subspaces. Instead, GFK has the advantage of being computationally efficient, automatically inferring the parameters without extensive cross-validation techniques [39], but has not been studied on remote sensing image classification. Moreover, the original GFK actually uses a linear kernel function, which may limit the performance in nonlinear feature transfer tasks. Thereby, the major objectives and contributions of this paper are:

(1) to propose and investigate geodesic Gaussian flow kernel SVM when hyperspectral image classification requires domain adaptation;

(2) to investigate the performances of different subspace feature extraction techniques, such as factor analysis (FA) [41], randomized nonlinear principal component analysis (rPCA) [42] and non-negative matrix factorization (NNMF) [43], with respect to the standard PCA, originally used in GFKSVM; and

(3) to compare the proposed methodology with conventional support vector machines (SVMs) and state-of-the-art DA algorithms, such as the information-theoretical learning of discriminative cluster for domain adaptation (ITLDC) [44], the joint distribution adaptation (JDA) [45], and the joint transfer matching (JTM) [46].

The rest of this paper is organized as follows. In Section 2, we formalize the framework used for unsupervised domain adaptation using the geodesic flow, and shortly recall the basics of the geodesic flow Gaussian kernel based support vector machines. Section 3 describes the datasets used and the setup of our experiments. Then, the experimental results are reported, analyzed and discussed for evaluation and comparison purposes in Section 4. Finally, in Section 5, we summarize the main findings of this paper and discuss several possible future research directions.

2. GFKSVM

2.1. Related Works

Let us consider that a specific domain D consists of a feature space X and the corresponding marginal distribution $P(X)$, where X = $\{x_1,x_2, \ldots ,x_n\} \in X$, so that $D = \{x, P(X)\}$. Furthermore, let us consider that a "task" Y is the joint set of a label space and a predictive function f, so that $\mathrm{Y} = \{y, f\}$. In general, if two domains are different, it means that either $\chi_S \neq \chi_T$ or $P_S(X) \neq P_T(X)$ or both. Similarly, the condition $\mathrm{Y}_S \neq \mathrm{Y}_T$ implies that either $y_S \neq y_T$ or $P\left(Y_S|X_S\right) \neq P\left(Y_T|X_T\right)$ or both, from a probabilistic point of view. In this scenario, TL aims at improving the learning of the predictive function f_{T} in the target domain D_T using the knowledge available in the source domain D_s and in the learning task Y_S when either $D_s \neq D_T$ or $\gamma_{\mathrm{S}} \neq \gamma_{\mathrm{T}}$.

As mentioned above, and according to the availability of source domain labeled data, supervised, semi-supervised or unsupervised feature construction methods can be adopted. Unsupervised domain adaptation is the only way when there is no correspondence between the domains to estimate the transformation between the labeled samples. In this area, dimensionality reduction via maximum mean discrepancy embedding (MMDE) and TCA were recently proposed [47,48]. The goal of unsupervised DA approached is to obtain meaningful intermediate subspaces corresponding to the D_s and D_T first, and then predict the target labels by exclusively using the labeled source training data. In this work, we focus on using GFK for unsupervised subspaces feature transfer in hyperspectral image classification, following the works in [38,39].

2.2. Geodesic Flow for DA

In a general metric space with distance metric, the length of a path between two points is defined as the smallest upper bound of any finite approximation of the path, a straight line in a Euclidean space (dashed line in Figure 1). However, the distance can be computed as a length (see the solid red line in Figure 1) in a geodesic metric space where Riemannian manifolds are used.

According to the theory of manifold learning (ML), the collection of all d-dimensional subspaces from the Grassmannian manifold $\mathbb{G}\left\{d,\mathfrak{D}\right\}$can be used to define a smooth Riemanian manifold with geometric, differential and probabilistic structures, where $\mathfrak{D}$ is the dimensionality of original data [38–40,49]. Since a full-fledged explanation for Grassmannian and Riemanian manifolds is beyond the scope of this paper, we refer the interested readers to papers [38–40,49,50]. Additionally, the d-dimensional subspaces in $\Re^{\mathfrak{D}}$ can be easily identified in a $\mathbb{G}\left\{d,\mathfrak{D}\right\}$. If two points mapped from the original subspaces D_s and D_T are closer in $\mathbb{G}\{d, \mathfrak{D}\}$, then the two domains could be made similar in this manifold, *i.e.*, $P_S(X) \cong P_T(X)$.

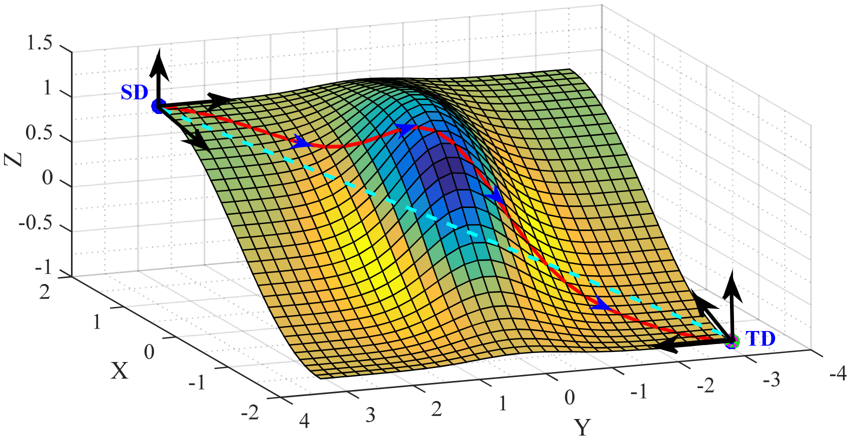


Figure 1. Illustration of the main idea of geodesic flow for domain adaptation.

Let $\mathcal{P}_S, \mathcal{P}_T \in \Re^{\mathfrak{D}\times\, d}$ denote the sets of orthogonal basis obtained by a principal component analysis (PCA) in the source and target domains. Moreover, let $\mathcal{R}_S \in \Re^{\mathfrak{D}\times(\mathfrak{D}-d)}$ represent the orthogonal complement to $\mathcal{P}_S$, *i.e.*, $\mathcal{R}_S^{\mathrm{T}^*}\mathcal{P}_S = 0$. In a Grassmannian manifold, using the canonical Euclidean metric for Riemannian manifold, with the constraints $\Phi(0) = \mathcal{P}_S$ and $\Phi(1) = \mathcal{P}_T$, the geodesic flow $\Phi : t \in [0, 1] \to \Phi(t) \in \mathbb{G}\{d, \mathfrak{D}\}$ from $\mathcal{P}_s$ to $\mathcal{P}_T$ can be formed as [39]:

$$\begin{aligned} \Phi(t) &= \mathcal{P}_s \boldsymbol{U}_1 \Gamma(t) - \mathcal{R}_S \boldsymbol{U}_2 \boldsymbol{\Sigma}(t) \\ &= \begin{bmatrix} \mathcal{P}_S & \mathcal{R}_S \end{bmatrix} \begin{bmatrix} \boldsymbol{U}_1 & 0 \\ 0 & \boldsymbol{U}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\Gamma}(t) \\ \boldsymbol{\Sigma}(t) \end{bmatrix} \end{aligned} \tag{1}$$

where $\boldsymbol{U}_1 \in \Re^{d\times d}$ and $\boldsymbol{U}_2 \in \Re^{\mathfrak{D}\times(\mathfrak{D}-d)}$ are the orthonormal matrices, given by the singular value decomposition (SVD):

$$\mathcal{P}_S^{\mathrm{T}^*}\mathcal{P}_T = \boldsymbol{U}_1 \boldsymbol{\Gamma} \boldsymbol{V}^{\mathrm{T}^*}, \mathcal{R}_S^{\mathrm{T}^*}\mathcal{P}_T = \boldsymbol{U}_2 \boldsymbol{\Sigma} \boldsymbol{V}^{\mathrm{T}^*} \tag{2}$$

where T* means matrix transpose, $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$ are diagonal matrices whose elements are cosθ_i and $sin\theta_i$, for, and θ_i are the angles measuring the degree of "overlap" between two subspaces $\mathcal{P}_s$ and $\mathcal{P}_T$.

2.3. Geodesic Flow Kernel SVM

According to Equation (1), the geodesic flow can be seen as a collection of infinite subspaces gradually varying from D_s to D_T. Starting from the original D-dimensional feature vector X, with projections denoted as $\Phi : t \in [0, 1]$, it is possible to form a feature set of infinite dimension $\boldsymbol{z}^\infty = \{\Phi(t)\,\boldsymbol{x} : t \in [0, 1]\}$. The inner product between two projected feature vectors with infinite dimensions $\boldsymbol{z}_i^\infty$ and $\boldsymbol{z}_j^\infty$ defines the GFK [39]:

$$\left\langle \boldsymbol{z}_i^\infty, \boldsymbol{z}_j^\infty \right\rangle = \int_0^1 \left(\boldsymbol{\Phi}(t)^{\mathrm{T}^*} \boldsymbol{x}_i \right)^{\mathrm{T}^*} \left(\Phi(t)^{\mathrm{T}^*} \boldsymbol{x}_j \right) dt = \boldsymbol{x}_i^{\mathrm{T}^*} \mathcal{G} \boldsymbol{x}_j \tag{3}$$

where $\boldsymbol{\mathcal{G}} \in \Re^{\mathfrak{D}\times\mathfrak{D}}$ is the positive semidefinite matrix

$$\boldsymbol{\mathcal{G}} = \int_0^1 \boldsymbol{\Phi}(t) \Phi(t)^{\mathrm{T}^*} dt \tag{4}$$

Substituting Equation (1) into Equation (4), and ignoring for the moment the constant part:

$$\mathcal{G} = \int_0^1 \begin{bmatrix} \Gamma(t)\Gamma(t) & -\Gamma(t)\Sigma(t) \\ -\Sigma(t)\Gamma(t) & \Sigma(t)\Sigma(t) \end{bmatrix} dt \tag{5}$$

where $\mathbf{\Gamma}(t)$ and $\mathbf{\Sigma}(t)$ are matrices whose elements are cos($t\theta_i$) and sin($t\theta_i$), respectively. Thus, by integrating in closed-form:

$$\lambda_{1,i} = \int_0^1 \cos^2(t\theta_i)\,dt = 1 + \frac{\sin(2\theta_i)}{2\theta_i} \tag{6}$$

$$\lambda_{2,i} = \int_0^1 \cos(t\theta_i)\sin(t\theta_i)\,dt = \frac{\cos(2\theta_i) - 1}{2\theta_i} \tag{7}$$

$$\lambda_{3,i} = \int_0^1 \sin^2(t\theta_i)\,dt = 1 - \frac{\sin(2\theta_i)}{2\theta_i} \tag{8}$$

where i = 1,2, ... ,d and, $\lambda_{1,i}$, $\lambda_{2,i}$ $\lambda_{3,i}$ become the i-th diagonal elements of the diagonal matrices $\mathbf{\Lambda}_1$, $\mathbf{\Lambda}_2$ and $\mathbf{\Lambda}_3$ respectively. Accordingly, matrix $\mathcal{G}$ assumes the form:

$$\mathcal{G} = \mathbf{\Omega}^{\mathrm{T}*} \begin{bmatrix} \mathbf{\Lambda}_1 & \mathbf{\Lambda}_2 \\ \mathbf{\Lambda}_3 & \mathbf{\Lambda}_4 \end{bmatrix} \mathbf{\Omega} \tag{9}$$

where $\mathbf{\Omega}$ represents the constant parts in Equation (1):

$$\mathbf{\Omega} = \begin{bmatrix} \mathcal{P}_S & \mathcal{R}_S \end{bmatrix} \begin{bmatrix} \mathrm{U}_1 & 0 \\ 0 & \mathrm{U}_2 \end{bmatrix} \tag{10}$$

It can be noted that Equation (3) is exactly the "kernel trick", where a *linear kernel* function induces inner products between infinite-dimensional features.

However, we may be faced with the situation of non-linear patterns in the data, patterns that linear kernel function cannot handle. Nonlinear kernel functions have proved instead to be effective in these situations not only for classification but also for feature extraction [51]. Hence, we apply to Equation (3) a non-linear transformation, and more specifically a kernel method [52]. A generic kernel $\boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is represented by:

$$\boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\gamma d_{\mathcal{G}}(\boldsymbol{x}_i, \boldsymbol{x}_j)^q\right), \gamma, q > 0 \tag{11}$$

Assuming γ = 1 and q = 2 the generic geodesic flow Gaussian kernel mapping function is obtained:

$$\boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left\{-d^2_{\mathcal{G}}(\boldsymbol{x}_i, \boldsymbol{x}_j)\right\} = \exp\left\{\frac{-(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}*}\,\mathcal{G}\,(\boldsymbol{x}_i - \boldsymbol{x}_j)}{\sigma^2}\right\} \tag{12}$$

where the parameter σ is a scaling factor.

Let now $\{X^S, Y^T\} = \{(x_i^S, y_i^S)\}_{i=1}^{n_S}, x_i^S \in \Re^{\mathfrak{D}}, y_i^S \in \{1, \ldots, M\}$ denote the set of n_s labeled source training data corresponding to M classes. $X^T = \left\{x_j^T\right\}_{j=1}^{n_T}, x_j^T \in \Re^{\mathfrak{D}}$ represents the set of n_T unlabeled data from D_T. Using a Lagrangian formulation to the classic linearly constrained optimization problem, the final dual problem can be rewritten as:

$$\begin{cases} \max\limits_{\alpha} \left\{ \sum\limits_{i=1}^{n_S} \alpha_i - \frac{1}{2}\sum\limits_{i=1}^{n_S}\sum\limits_{j=1}^{n_S} \alpha_i\alpha_j y_i^S y_j^S \boldsymbol{K}\left(\boldsymbol{x}_i^S, \boldsymbol{x}_j^S\right)\right\} \\ s.t.: \sum\limits_{i=1}^{n_S} y_i^S\alpha_i, 0 \leqslant \alpha_i, \forall i = 1, \ldots, n_S \end{cases} \tag{13}$$

where the kernel matrix $\boldsymbol{K}\left(\boldsymbol{x}_i^S, \boldsymbol{x}_j^S\right)$ is computed either by Equation (3) or Equation (12). Thereby, the decision function for any test vector $\boldsymbol{x}_j^T$ from X^T is finally given by:

$$f\left(\boldsymbol{x}_*^T\right) = \operatorname{sgn}\left(\sum_{i=1}^{N^{SVs}} y_i^S \alpha_i \boldsymbol{K}\left(\boldsymbol{x}_i^S, \boldsymbol{x}_*^T\right) + b\right) \tag{14}$$

where α_i are the Lagrange multipliers, N^{SVS} is the total numbers of support vectors (SVs), and the bias term b can be obtained by using the k unbounded Lagrange multipliers according to:

$$b = \frac{1}{k\sum_{i=1}^{k}\left(y_i^S - \left\langle \Phi\left(\boldsymbol{x}_i^S\right), \boldsymbol{w}\right\rangle\right)} \tag{15}$$

$$\mathbf{w} = \sum_{i=1}^{N^{SVs}} y_i^S \alpha_i \Phi\left(\boldsymbol{x}_i^S\right) \tag{16}$$

where $\Phi(\cdot)$ denotes the geodesic projected feature vectors.

In summary, the steps of the GFKSVM algorithm may be reported as:

```
Algorithm 1: GFKSVM
Inputs:  Source domain image $S_{img}$; Training samples from source domain: $X_s$; Target domain
         image $T_{img}$; Validation samples from target domain: $X_T$; Subspace construction approach
         for $S_{img}$: $FE_S \in \{PCA, rPCA, FA, NNMF\}$; Subspace construction approach for $T_{img}$: $FE_T \in$
         $\{PCA, rPCA, FA, NNMF\}$; number of transferring subspaces $d$; Kernel type: $K$; Geodesic
         flow kernel function: $\boldsymbol{\mathcal{G}}$.
Train:
    (1) Extract features able to represent the source domain sunspaces: $S_{img}^{sub} = FE_s(S_{img})$
        and $T_{img}^{sub} = FE_T(T_{img})$;
    (2) Compute the geodesic flow kernel $\boldsymbol{\mathcal{G}} = G(S_{img}^{sub}, T_{img}^{sub}, d)$ according to Equation (9);
    (3) Compute the kernel matrix $\mathrm{K}_{\mathrm{X_s}}$ for $X_s$ by means of Equation (3) or Equation (12)
        according to the Kernel type $K$;
    (4) Compute the SVM model parameters by solving Equations (13);
    (5) Compute the kernel matrix $\mathrm{K}_{\mathrm{XT}}$ for $\mathrm{X_T}$ by Equation (3) or Equation (12) according to the
        Kernel type $K$;
Classify:
    Return the predicted label for $\mathrm{X_T}$ according to Equation (14).
```

3. Datasets and Setup

3.1. Datasets Descriptions

For the experimental analysis, two hyperspectral datasets with different degrees of shift between source and target domains were considered.

For the first test case, two hyperspectral images collected by the Reflective Optics Spectrographic Image System (ROSIS) sensor over the University of Pavia and Pavia City Centre were considered (see Figure 2). The ROSIS optical sensor provides up to 115 bands with a spectral range coverage ranging from 0.43 µm to 0.86 µm, and a spatial resolution of 1.3 meters per pixel. The Pavia City Centre image contains 102 spectral bands and has a size of 1096 × 1096 pixels. The Pavia University image contains instead 103 spectral reflectance bands and has a size of 610 × 3400 pixels. Both images are provided with ground truths of nine classes for each. Seven classes are shared by both images, and were considered in our experiments (Table 1). In the experiments, the Pavia University image was considered as the source domain, while the Pavia City Center image as the target domain, or vice versa. Differences in imaging conditions, roof materials, and vegetation types cause remarkable variations of spectral and statistical properties of these land-cover classes across the images, as observed in Figure 2. Moreover, since the GFKSVM is not suitable for situations where the source and target domains are represented by a different number of features, only 102 spectral bands of the Pavia University image were used in the experiments.

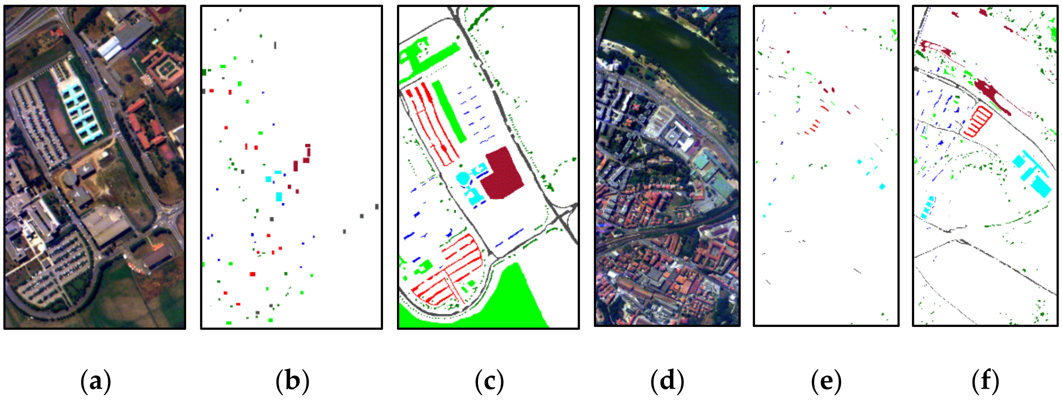


Figure 2. Color composite images of the (**a**) ROSIS Pavia University with the corresponding ground truths: (**b**) train map; (**c**) test map; (**d**) Pavia City Centre with the corresponding ground truths; (**e**) train map and (**f**) test map.

Table 1. Class legend for ROSIS University and City Center datasets.

No.	Class	Code	University		Center	
			Train	Test	Train	Test
1	Asphalt	⬛	548	6631	678	7585
2	Meadows	🟩	540	18649	797	2905
3	Trees	🟩	524	3064	785	6508
4	Bare soil	🟫	532	5029	820	6549
5	Bricks	🟥	514	3682	485	2140
6	Bitumen	🟦	375	1330	808	7287
7	Shadows	🟦	231	947	195	2165

The second dataset is a 2.5 spatial resolution hyperspectral image consisting of 144 spectral bands in the 0.38–1.05 µm region. The data were acquired by the National Science Foundation (NSF)-funded Center for Airborne Laser Mapping (NCALM) over the University of Houston campus and the neighboring urban area on 3 June 2012, it is freely provided by Hyperspectral Image Analysis Lab affiliated with the University of Houston's Electrical and Computer Engineering Department. Originally, the data sets have a size of 1905 × 349 pixels and their ground truth includes 15 land cover types. Similarly to the previous case, we considered two disjoint sub-images with 750 × 349 pixels (Figure 3a) and 1155 × 349 pixels (Figure 3b), respectively. These sub-images share eight classes in the ground truth: healthy grass, stressed grass, trees, soil, residential, commercial, road and parking lots, listed in Table 2 with the corresponding sample size.

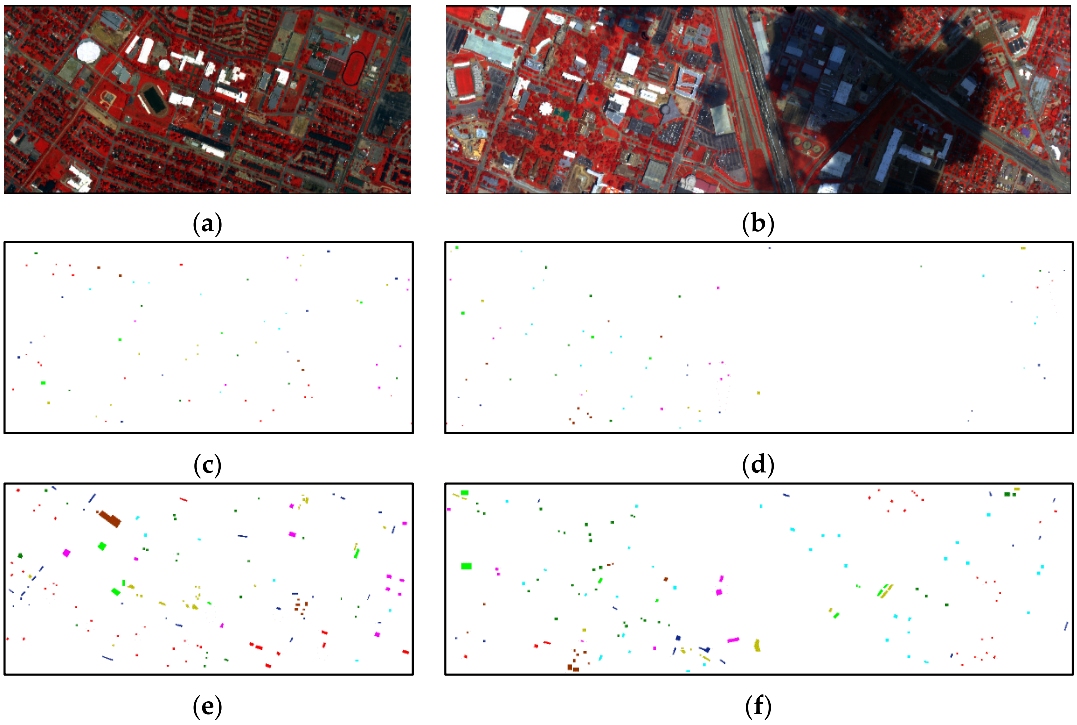
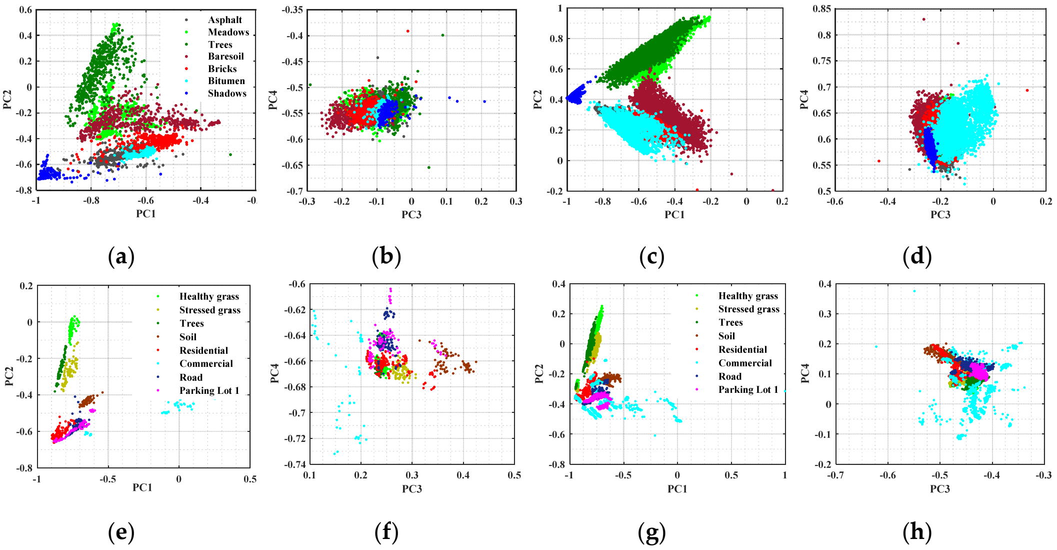

(**a**) (**b**)
(**c**) (**d**)
(**e**) (**f**)

Figure 3. Color composite of the Source (**a**) and Target (**b**) images of Houston with the corresponding ground truths: (**c**) and (**d**) train maps for Source and Target; (**e**) and (**f**) test maps for Source and Target.

Table 2. Class legend for Figure 3.

No.	Class	Code	Left		Right	
			Train	Test	Train	Test
1	Healthy grass	🟩	98	449	100	604
2	Stressed grass	🟨	87	482	103	582
3	Trees	🟩	78	373	110	683
4	Soil	🟥	72	688	114	368
5	Residential	🟥	173	687	23	385
6	Commercial	🟦	47	132	144	921
7	Road	🟦	108	589	85	470
8	Parking Lot 1	🟪	88	625	104	416

To provide a better illustration of the differences between the statistical properties of the same classes in two domains, Figure 4 presents the principal component distributions of different classes from the source and target images for the Pavia and Huston data sets. Clear distribution shifts for Shadows, Bitumen, Asphalt, Bricks and Bare soil can be observed in Figure 4a,c for Pavia. Similarly, please note the large distribution shifts for Soil, Commercial and Parking lot 1 between the two domains in Figure 4e,g for Houston.



(**a**) (**b**) (**c**) (**d**)

(**e**) (**f**) (**g**) (**h**)

Figure 4. Principal component scatter plots of the source and target images for Pavia (University: (**a**) PC1 *vs.* PC2; (**b**) PC3 *vs.* PC4; Center: (**c**) PC1 *vs.* PC2; (**d**) PC3 *vs.* PC4 and Houston Left: (**e**) PC1 *vs.* PC2; (**f**) PC3 *vs.* PC4; Right: (**g**) PC1 *vs.* PC2; (**h**) PC3 *vs.* PC4.) data sets.

3.2. Experimental Setup

In the experiments, we compare the results of SVMs with radial bias function (RBF) kernel applied to the following features: "All bands"—the original spectral bands; "$(\mathrm{S})_{\mathrm{pca}}$–$(\mathrm{T})_{\mathrm{pca}}$"–difference of PCA/FA/NMF/rPCA transformations separately applied to the source and target images; "$(\mathrm{S} + \mathrm{T})_{\mathrm{pca}}$"—PCA/FA/NMF/rPCA transformations applied to a combination of the source and target images; "GFK"—unsupervised subspace features transferred from geodesic flow kernel. Free SVM parameters such as γ and C were tuned in the range (0.01–10) and (1–100), respectively, by a grid searching criterion, and 10-fold cross validation.

As mentioned in the introduction, we also compare results using the ITLDC, JDA and JTM algorithms. As for ITLDC and JDA, the free parameter λ was tuned in the range (0.01–10) with the three cross validations technique. In JTM, an RBF kernel function was applied to build the joint kernel matrix, and the tuning range for the γ factor of this kernel, as well as the λ parameter used for eigenvalue decomposition, were selected in the (0.01–10) range. For a more objective comparison, SVM with RBF kernel was used in the target domain classification phase in ITLDC, JDA and JTM.

All the experiments were carried out using Matlab™ on a Windows 7 64 bit operating system with an x64-based processor Intel® Core™ i7-4790 CPU, @3.60 GHz, 32 GB RAM.

4. Experimental Results

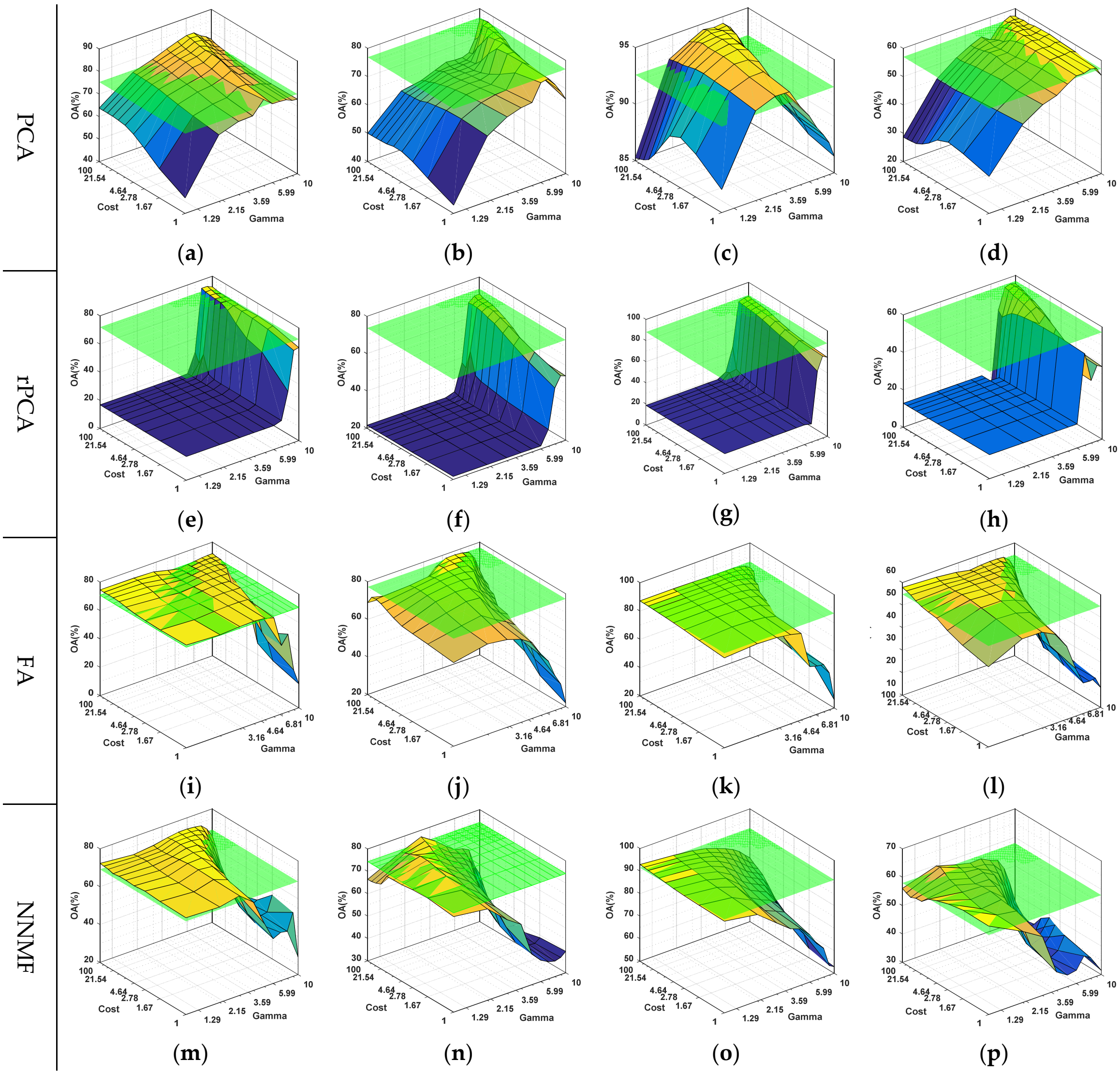
4.1. Parameter Evaluation

4.1.1. Kernel Parameter Evaluation for GFKSVM

Generally, it is expected that an RBF kernel is more suitable than a linear kernel in case of a nonlinear classification task. However, the free tuned parameter of the RBF kernel in the source domain may not be optimal for the target domain, hence limiting or even decreasing the performance of GFKSVM. While it has been proved that GFKSVM has excellent performances with respect to a linear kernel in computer vision adaptation tasks [38,39], its performance on real hyperspectral data with RBF kernel remains to be tested. Hence, we comparatively investigated the performance of GFKSVM with RBF and linear kernels first.

Figure 5 presents the Overall Accuracy (OA) values for GFKSVM with RBF and linear kernel applied to the ROSIS data with different subspace feature transfer approaches, but always with

10 transferred features. The selection of 10 features is completely arbitrary, as here we focus on the comparison of linear and non-linear kernels. According to the graphs in Figure 5, GFKSVM with RBF kernel shows higher OA values than the linear kernel in most cases, when train and test are carried out only in the source domain (see Figure 5a,c,e,i,m)). However, this advantage decreases when the training and test are carried out in the source and target domains separately (see Figure 5b,d,f,j,n). These results tell us that the free parameters for RBF kernel are capable of reaching higher OA values than the linear kernel.



(**a**) (**b**) (**c**) (**d**)

(**e**) (**f**) (**g**) (**h**)

(**i**) (**j**) (**k**) (**l**)

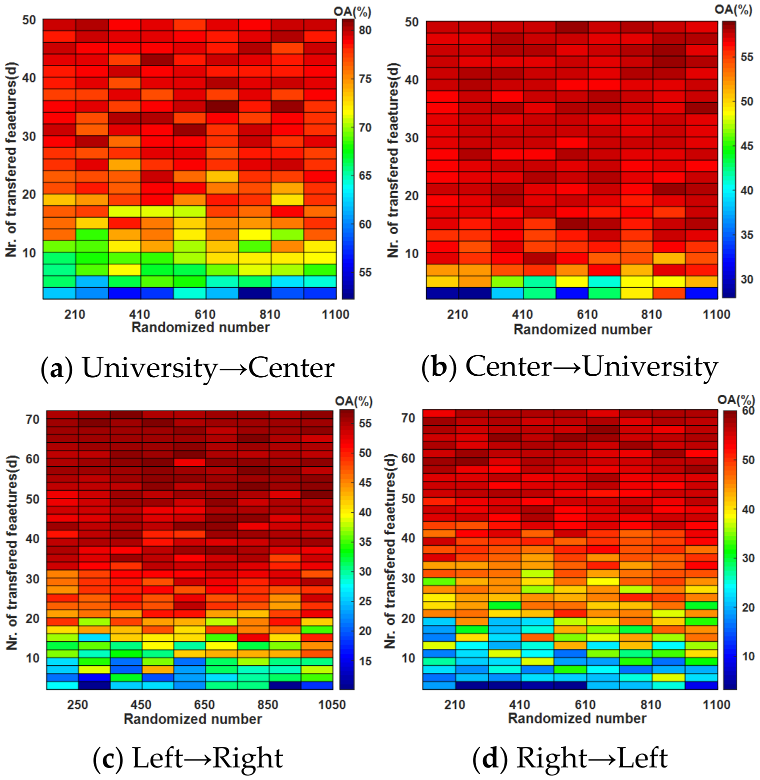
(**m**) (**n**) (**o**) (**p**)

Figure 5. Overall Accuracy (OA) *versus* GFKSVM with RBF and linear kernel (green transparent flat surface) for ROSIS data (PCA-PCA: (**a**) University→University; (**b**) University→Center; (**c**) Center→ Center; (**d**) Center→University; rPCA-rPCA: (**e**) University→University; (**f**) University→Center; (**g**) Center→Center; (**h**) Center→University; FA-FA: (**i**) University→University; (**j**) University→Center; (**k**) Center→Center; (**l**) Center→University; NNMF-NNMF: (**m**) University→University; (**n**) University→ Center; (**o**) Center→Center; (**p**) Center→University), with number of subspaces $d = 10$.

4.1.2. Parameter Evaluation for rPCA

Randomized nonlinear principal component analysis (rPCA) is a nonlinear variant of PCA recently proposed as a low-rank approximation of kernel principal component analysis (KPCA) [42]. In comparison with PCA, rPCA is capable of revealing nonlinear patterns in data with computational complexity of $O(m^2n)$, as opposed to a computational complexity of $O(n^3)$ for KPCA and $O(d^2n)$ for PCA, where m, n and d represent the number of random features, the sample size and the number of extracted features, respectively, and $m \geqslant d$. According to [42], rPCA performances are mainly controlled by the number of random features (m) and extracted features (d). In the following, we investigate the performances of GFKSVM using rPCA for feature transfer. Figure 6 shows the OA results as a function of the parameters m and d.

From the results in Figure 6, it can easily observed that, in general, the larger the number of extracted features d, the higher the OA values. Instead, the role of the number of random features m is less definite, although it is clear it should be a large value. Accordingly, this parameter was set to five times the original dimensionality of data in all the experiments.



(**a**) University→Center (**b**) Center→University

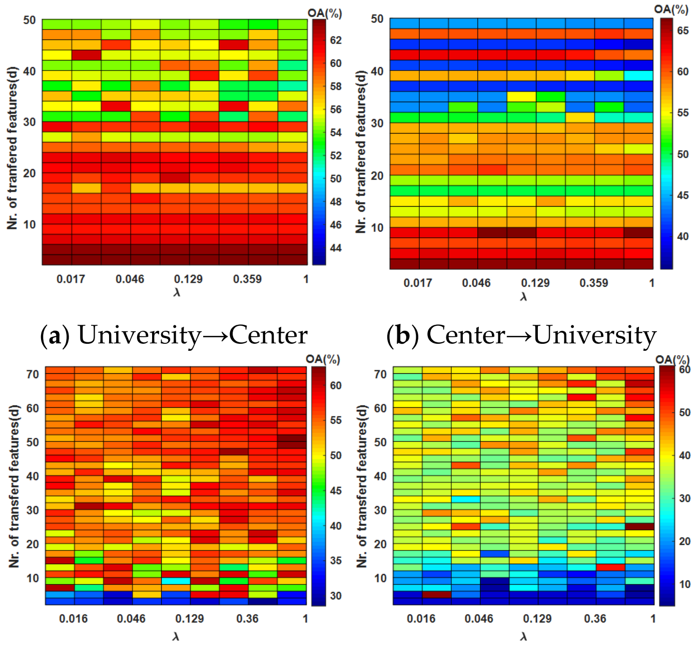
(**c**) Left→Right (**d**) Right→Left

Figure 6. OA *versus* the parameters m and d of rPCA in GFKSVM for ROSIS (**a**,**b**) and Houston (**c**,**d**) data sets: (Source→Target).

4.1.3. Parameter Evaluation for ITLDC

Additionally, performance analysis using ITDLC was conducted. In order to understand the results, please recall that ITLDC identifies a feature space where data from source and target domains are similarly distributed [43]. Its performances are controlled by the dimensionality of the transferred feature subspace d and the regularization coefficient λ. Similarly to previous experiments, Figure 7 depicts the OA values using ITLDC as a function of its free parameters.

According to the results in Figure 7, the regularization coefficient does not show significant influence on the OA values. Instead, the effects of the transferred feature subspace (d) are quite obvious and different for different data sets. Specifically, the optimal transferred feature subspace (d) for the relatively heterogeneous ROSIS data is $\leqslant$26 for University (Source)→Center (Target) and $\leqslant$10 for Center (Source)→University (Target) DA learning. Instead, for the more homogeneous Houston data set, the larger the transferred feature subspace (d), the higher the OA value.



(**a**) University→Center (**b**) Center→University

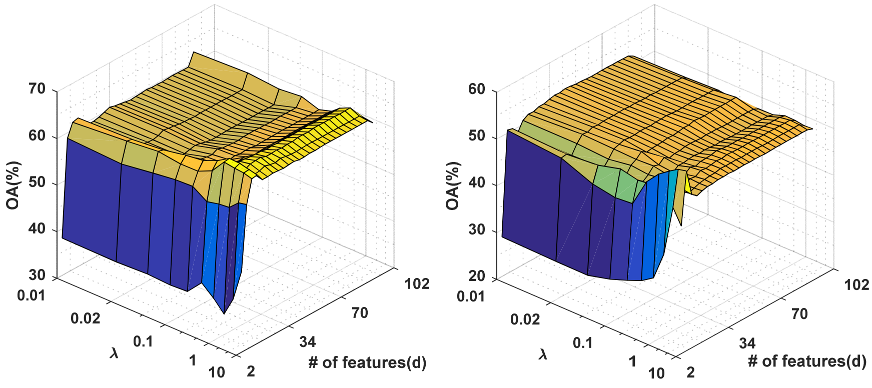
(**c**) Left→Right (**d**) Right→Left

Figure 7. OA *versus* the number of subspaces *d* and the parameter λ of ITLDC for the ROSIS (**a**,**b**) and Houston (**c**,**d**) data sets: (Source→Target).

4.1.4. Parameter Evaluation for JDA

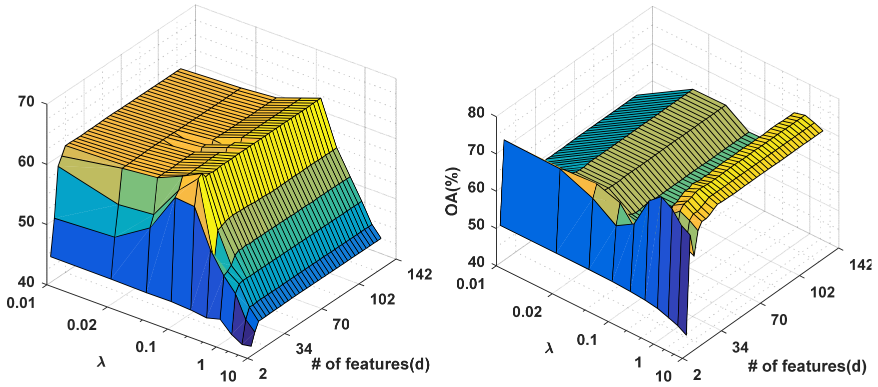
The joint distribution adaptation (JDA) algorithm simultaneously reduces the difference in both the marginal distribution and conditional distribution between domains [45]. Specifically, JDA uses an extended nonparametric maximum mean discrepancy MMD [53] to measure the differences in both marginal and conditional distributions first, and then integrates it with PCA for constant effective and robust feature representation. According to [45], the performance of JDA is controlled by the regularization parameter λ and the transferred feature subspace d. JDA has a high computational cost $O(TdD^2 + TCn^2 + TDn)$, where T is the iteration number, C represents the cardinality label, and D denotes the original data dimensionality.

According to the results in Figure 8, a larger transferred feature number ($d \geqslant 10$) leads to higher and stable OA values, but the main control power is retained by the regularization parameter λ. Indeed, with a constant and large set of transferred features, only the regularization parameter needs to be tuned.



(**a**) University→Center (**b**) Center→University

Figure 8. *Cont.*



(**c**) Left→Right (**d**) Right→Left

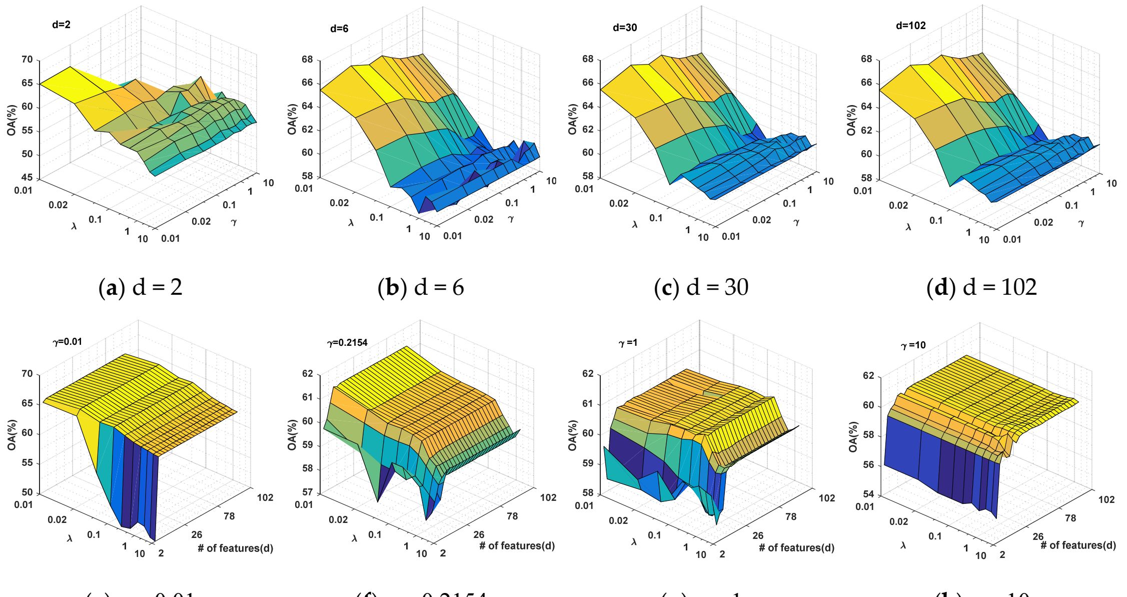
Figure 8. OA *versus* the number of subspaces *d* and the parameter λ of JDA (trained with primal) for the ROSIS (**a**,**b**) and Houston (**c**,**d**) data sets: (Source→Target).

4.1.5. Parameter Evaluation for JTM

Joint transfer matching (JTM) aims at reducing the domain shift by jointly matching the features and reweighting the instances in a principal dimensionality reduction procedure [46]. Briefly, the feature matching procedure is executed by minimizing the nonparametric MMD [53] in an infinite dimension reproducing kernel Hilbert space (RKHS), while the instance reweighting and domain invariant feature representation is performed by minimizing the $\ell_{2,1}$-norm using PCA. The overall algorithmic computational complexity for JTM is $O(TdD^2 + TCn^2)$, and free parameters are the regularization coefficient λ, the factor γ for RBF kernel, and the transferred feature subspace *d*. Figure 9 shows the OA values as a function of these free parameters.

It can be observed that with a constant and large set of the transferred features ($d \geqslant 6$), the performance of JTM for ROSIS data is mainly controlled by the regularization coefficient λ and the factor γ (Figure 9b–d).

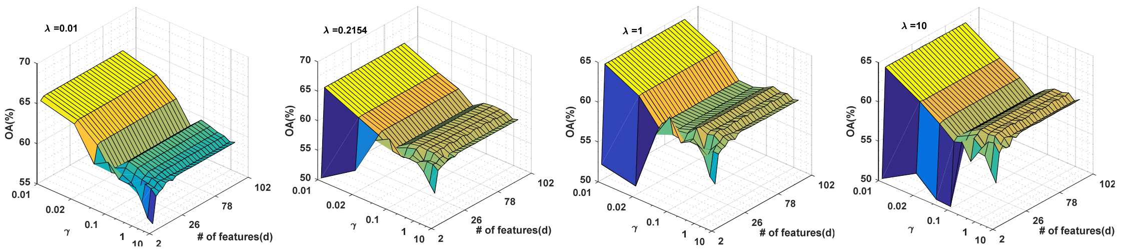
According to the results in Figure 10, once again the larger numbers of transferred features ($d \geqslant 6$, see Figure 10b–d), and smaller values for both the regularization coefficient (λ ~ 0.01) and the gamma factor (γ ~ 0.01) (Figure 10b–e,i) lead to the most effective results for the Houston data.



(**a**) d = 2 (**b**) d = 6 (**c**) d = 30 (**d**) d = 102

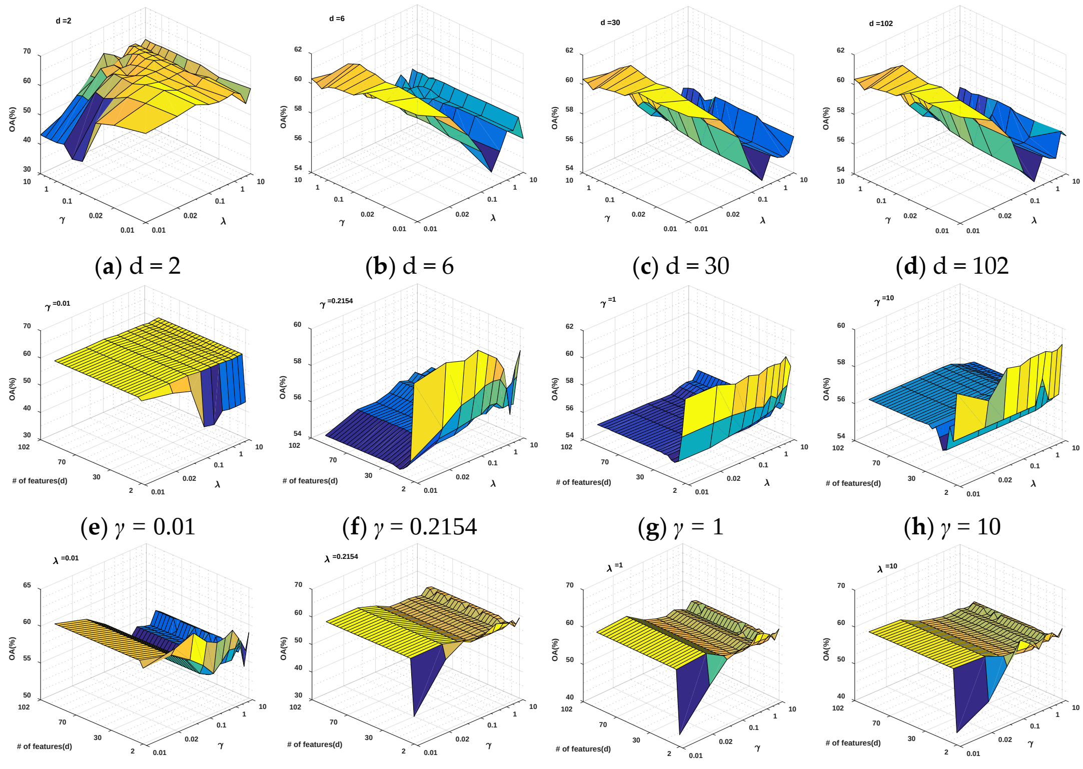
(**e**) γ = 0.01 (**f**) γ = 0.2154 (**g**) γ = 1 (**h**) γ = 10

Figure 9. *Cont.*



(**i**) λ = 0.01 (**j**) λ = 0.2154 (**k**) λ = 1 (**l**) λ = 10

Figure 9. OA *versus* the number of features d, the SVM parameter γ, and the regularization factor λ of JTM for ROSIS data sets (University→Center). (**a**) OA *versus* γ and λ, d = 2; (**b**) OA *versus* γ and λ, d = 6; (**c**) OA *versus* γ and λ, d = 10; (**d**) OA *versus* γ and λ, d = 102; (**e**) OA *versus* λ and d, γ = 0.01; (**f**) OA *versus* λ and d, γ = 0.2154; (**g**) OA *versus* λ and d, γ = 1; (**h**) OA *versus* λ and d, γ = 10; (**i**) OA *versus* γ and d, λ = 0.01; (**j**) OA *versus* γ and d, λ = 0.2154; (**k**) OA *versus* γ and d, λ = 1; (**l**) OA *versus* γ and d, λ = 10.



(**a**) d = 2 (**b**) d = 6 (**c**) d = 30 (**d**) d = 102

(**e**) γ = 0.01 (**f**) γ = 0.2154 (**g**) γ = 1 (**h**) γ = 10

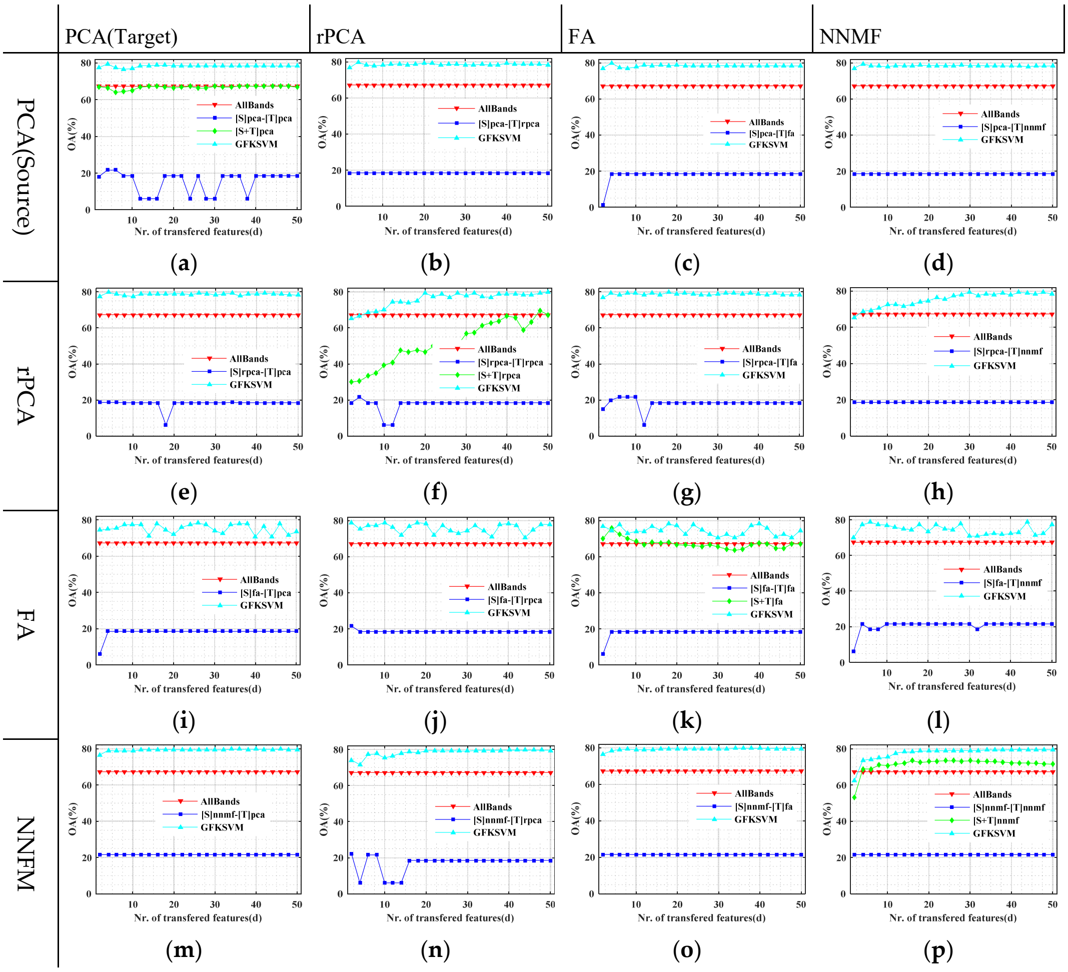
(**i**) λ = 0.01 (**j**) λ = 0.2154 (**k**) λ = 1 (**l**) λ = 10

Figure 10. OA *versus* the number of features *d*, the SVM parameter γ, and the regularization factor λ of JTM for ROSIS data sets (Left portion→Right portion). (**a**) OA *versus* γ and λ, d = 2; (**b**) OA *versus* γ and λ, d = 6; (**c**) OA *versus* γ and λ, d = 10; (**d**) OA *versus* γ and λ, d = 102; (**e**) OA *versus* λ and d, γ = 0.01; (**f**) OA *versus* λ and d, γ = 0.2154; (**g**) OA *versus* λ and d, γ = 1; (**h**) OA *versus* λ and d, γ = 10; (**i**) OA *versus* γ and d, λ = 0.01; (**j**) OA *versus* γ and d, λ = 0.2154; (**k**) OA *versus* γ and d, λ = 1; (**l**) OA *versus* γ and d, λ = 10.

4.2. Evaluation of GFKSVM with Different Feature Transfer Approaches

To complete our analysis, in Figures 11–13 we report the results of a sensitivity analysis of GFK with respect to its critical parameters: the number of transferred features *d* and the adopted unsupervised feature transfer approach (PCA, rPCA, FA and NNMF). According to Equations (1) and (2), the number of transferred subspaces (d) was set to $\mathfrak{D}$. Hence, the maximum d for ROSIS and Houston in Figures 11–13 was set to 50 and 72, respectively. Moreover, according to the results shown in the previous subsection, the number of random features m in rPCA is set to five times the dimensionality of the original data. Finally, all results are compared with the results obtained by directly using RBF kernel based SVM on the original data. These results are considered as the evaluation benchmark.

As illustrated in Figure 11 for the ROSIS dataset (University→Centre), the largest overall accuracy values are achieved by GFKSVM, while the second best performance is obtained in most cases by using all bands and RBF kernel based SVM. Furthermore, the sensitivity of GFK with respect to the subspace dimensionality is much lower for conventional methods than applying PCA, FA and NNMF transformations to the source and target images in either a separate or combined way. The optimum d for rPCA and NNMF based GFKSVM is $\geqslant$20 (see Figure 11f,p), while PCA based GFKSVM is much more stable with respect to the number of transferred features.



(**a**) (**b**) (**c**) (**d**)

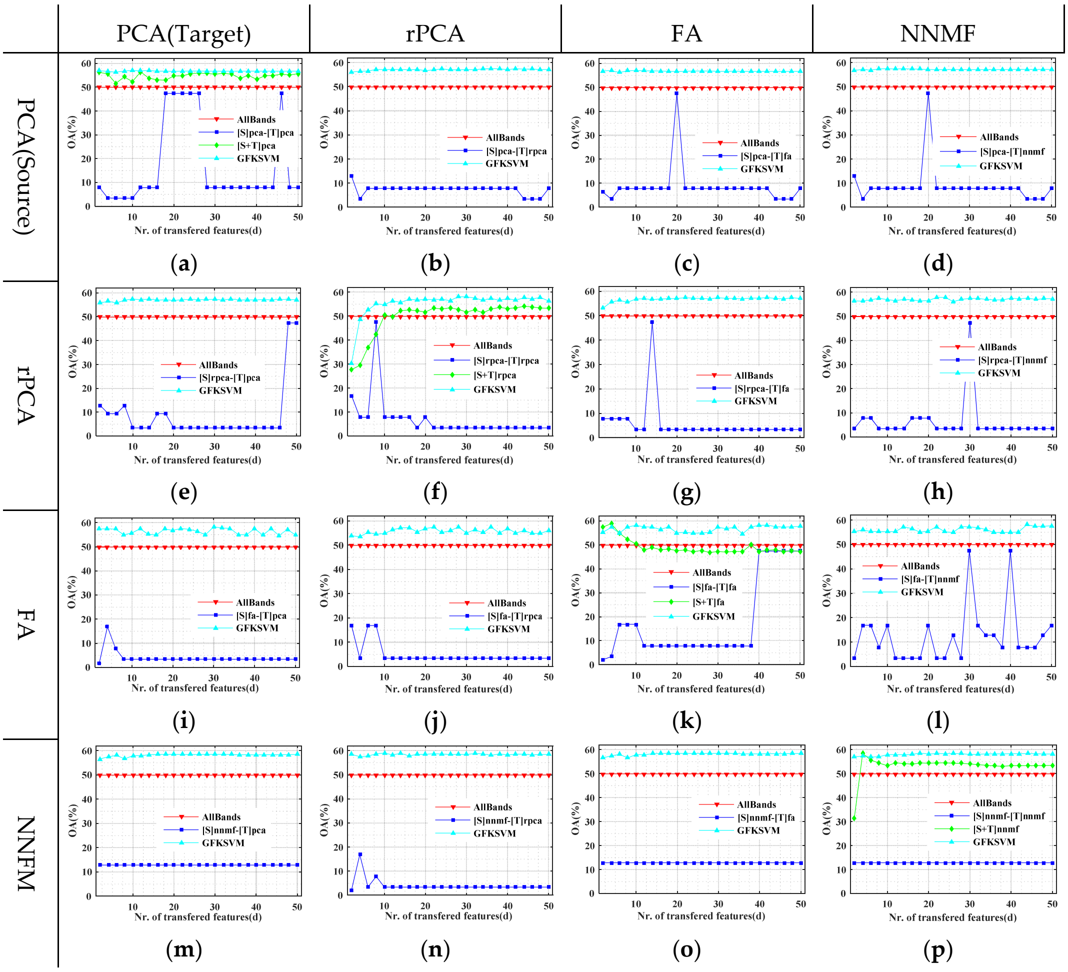
(**e**) (**f**) (**g**) (**h**)

(**i**) (**j**) (**k**) (**l**)

(**m**) (**n**) (**o**) (**p**)

Figure 11. Classification accuracy curves for GFKSVM using different subspace feature transfer approaches on the ROSIS data set (University→Center). (**a**) PCA-PCA; (**b**) PCA-rPCA; (**c**) PCA-FA; (**d**) PCA-NNMF; (**e**) rPCA-PCA; (**f**) rPCA-rPCA; (**g**) rPCA-FA; (**h**) rPCA-NNFM; (**i**) FA-PCA; (**j**) FA-rPCA; (**k**) FA-FA; (**l**) FA-NNMF; (**m**) NNMF-PCA; (**n**) NNMF-rPCA; (**o**) NNMF-FA; (**p**) NNMF-NNMF.

Figure 12 presents the results for GFKSVM with different subspace feature transfer approaches for Pavia, and Figure 13 presents that for Houston. Once again, all these results show that PCA based GFKSVM is more robust with respect to the dimensionality of the transferred features (d) than rPCA, FA and NMF approaches (Figures 12a,f,k,p and 13a,f,k,p). Indeed, rPCA and NMF require a larger *d* (Figures 12f,p and 13f,p). Instead, the poor performance of GFKSVM if rPCA is applied to both domains with small *d* can be significantly improved by replacing rPCA with PCA, FA or NNMF for one domain (see for instance Figure 11f *vs.* Figure 11e,g,h), Figure 12f *vs.* Figure 11e,g,h or Figure 13f *vs.* Figure 11e,g,h.



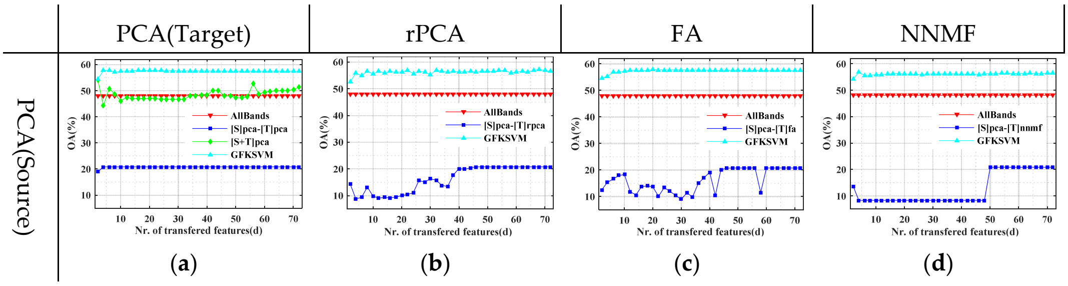
(**a**) (**b**) (**c**) (**d**)

(**e**) (**f**) (**g**) (**h**)

(**i**) (**j**) (**k**) (**l**)

(**m**) (**n**) (**o**) (**p**)

Figure 12. Classification accuracy curves for GFKSVM with various subspace feature transfer approaches on ROSIS data (Center→University). (**a**) PCA-PCA; (**b**) PCA-rPCA; (**c**) PCA-FA; (**d**) PCA-NNMF; (**e**) rPCA-PCA; (**f**) rPCA-rPCA; (**g**) rPCA-FA; (**h**) rPCA-NNFM; (**i**) FA-PCA; (**j**) FA-rPCA; (**k**) FA-FA; (**l**) FA-NNMF; (**m**) NNMF-PCA; (**n**) NNMF-rPCA; (**o**) NNMF-FA; (**p**) NNMF-NNMF.



(**a**) (**b**) (**c**) (**d**)

Figure 13. *Cont.*

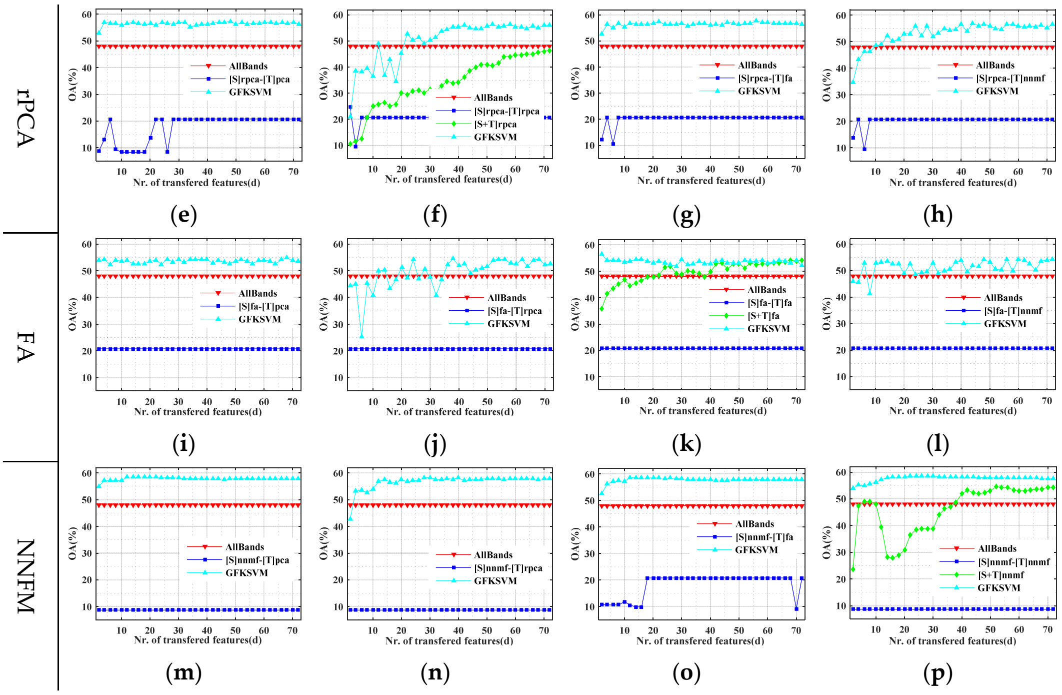


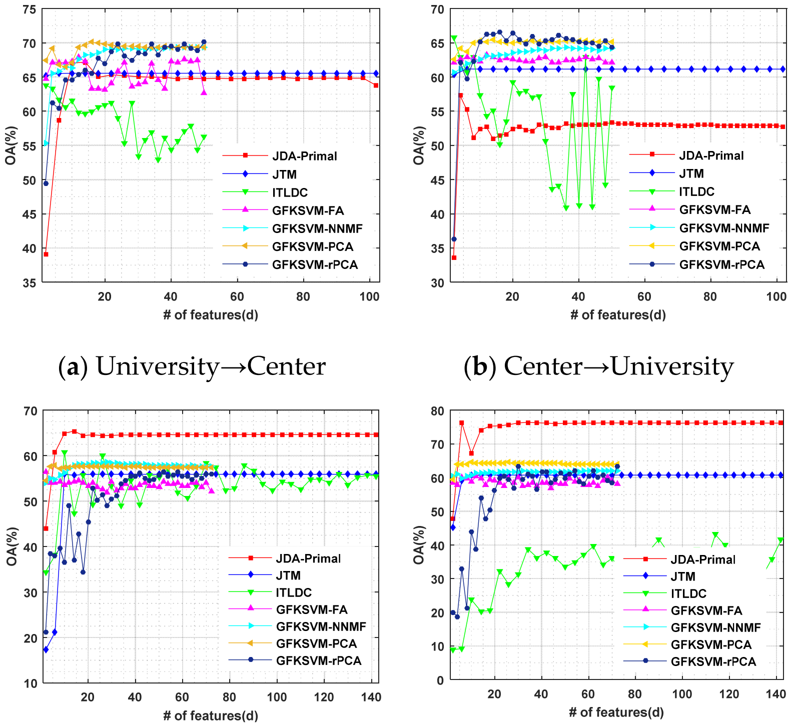
Figure 13. Classification accuracy curves for GFKSVM with various subspace feature transfer approaches on Houston data (Left site→Right site). (**a**) PCA-PCA; (**b**) PCA-rPCA; (**c**) PCA-FA; (**d**) PCA-NNMF; (**e**) rPCA-PCA; (**f**) rPCA-rPCA; (**g**) rPCA-FA; (**h**) rPCA-NNFM; (**i**) FA-PCA; (**j**) FA-rPCA; (**k**) FA-FA; (**l**) FA-NNMF; (**m**) NNMF-PCA; (**n**) NNMF-rPCA; (**o**) NNMF-FA; (**p**) NNMF-NNMF.

4.3. GFKSVM vs. State-of-the-Art DA Algorithms

Another set of experiments was devoted to the comparison of GFKSVM with ITLDC, JDA and JTM. In consideration of the high computational complexity of the latter algorithms, the training samples for the Pavia Center data set were used to validate the algorithms trained with training samples for the Pavia University data set, and vice versa. Figure 14 reports the OA curves as a function of the *d* parameter for all algorithms. Note that GFKSVM uses the same subspace feature reconstruction approach (PCA/FA/NMF/rPCA) for both the source and the target domain.

According to the results in Figure 14, GFKSVM after PCA, rPCA or NNMF shows in general larger OA values with respect to ITLDC, JDA and JTM. Instead, for the Houston data set, Figure 14c,d, JDA achieves the largest OA value. We may summarize that GFKSVM is more suitable for more statistically heterogeneous data like the ROSIS data set, while JDA is preferred for statistically homogeneous data like the Houston data set. Looking at this figure, we can also state that ITLDC is the best choice when a small number of features for heterogeneous data or a large number of features for homogeneous data are considered.

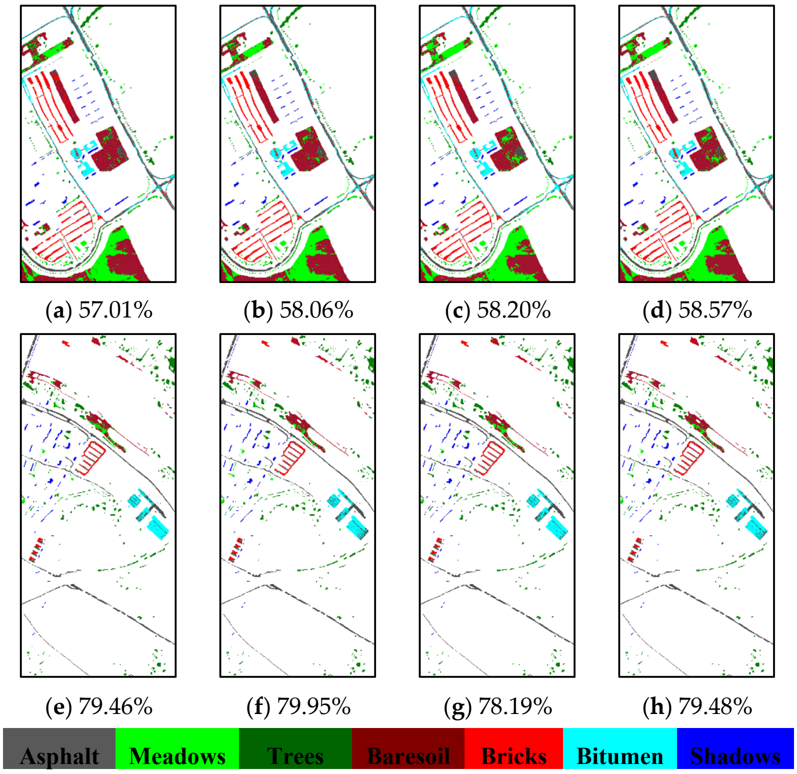
For a visual comparison, Tables 3 and 4 report the user accuracy (UA), kappa (κ) and average accuracy (AA) for the best result of every transformation, while Figure 15 presents the classification maps by GFKSVM with different subspace feature construction approaches for the ROSIS dataset. Please note that, since the validation procedures for JDA, ITLDC and JTM are carried out using the training samples for the Center image, while they are trained with the training samples for the University image, or vice versa, thematic maps are not presented.



(**a**) University→Center (**b**) Center→University

(**c**) Left→Right (**d**) Right→Left

Figure 14. OA curves for GFKSVM and other considered state-of-the-art DA methods on ROSIS (**a**,**b**) and Houston (**c**,**d**) data sets.



(**a**) 57.01% (**b**) 58.06% (**c**) 58.20% (**d**) 58.57%

(**e**) 79.46% (**f**) 79.95% (**g**) 78.19% (**h**) 79.48%

Figure 15. Classification maps (limited to validation/training sample locations) for GFKSVM on ROSIS data: Center→University: (**a**) PCA-PCA; (**b**) rPCA-rPCA; (**c**)-FA-FA; (**d**) NNMF-NNMF; University→Center: (**e**) PCA-PCA; (**f**) rPCA-rPCA; (**g**)-FA-FA; (**h**) NNMF-NNMF.

Table 3. Classification accuracies (%) and kappa statistic (k) for considered DA algorithms applied to ROSIS data (University→Center).

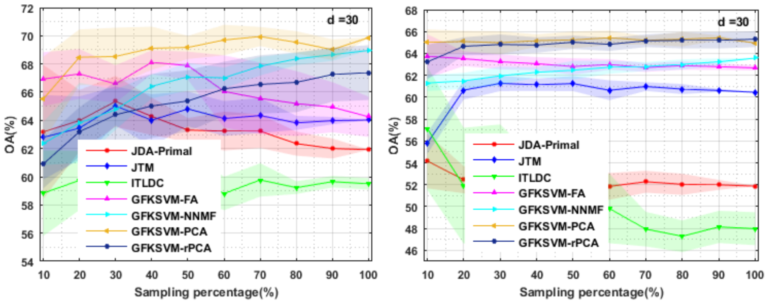
Class	GFKSVM				JDA	ITLDC	JTM
	PCA	rPCA	FA	NNMF			
Asphalt	97.73	**98.05**	97.82	98.04	90.27	95.87	85.55
Meadows	21.45	25.13	7.37	19.72	**47.30**	17.31	9.54
Trees	97.51	97.66	99.08	99.15	96.69	92.74	**99.49**
Bare soil	76.62	81.68	75.08	79.84	85.00	**88.29**	74.51
Bricks	75.65	**75.98**	75.65	75.14	56.49	63.92	61.03
Bitumen	65.02	60.85	64.43	61.30	**71.04**	36.51	62.75
Shadows	99.91	99.91	99.86	99.91	64.10	**100.00**	82.05
AA	76.27	77.04	74.18	76.16	72.98	70.66	67.85
OA	79.46	**79.95**	78.19	79.48	74.82	66.55	65.92
κ	0.75	0.76	0.74	**0.75**	0.70	0.60	0.60

Table 4. Classification accuracies (%) and kappa statistic (k) for considered DA algorithms applied to Houston data (Right→Left).

Class	GFKSVM				JDA	ITLDC	JTM
	PCA	rPCA	FA	NNMF			
Healthy grass	55.00	62.00	51.00	76.00	**100.00**	83.07	99.55
Stressed grass	**100.00**	**100.00**	**100.00**	**100.00**	89.83	62.66	93.15
Trees	**100.00**	99.09	96.36	**100.00**	99.73	93.57	0.54
Soil	70.18	72.81	71.05	71.93	**99.71**	77.33	98.98
Residential	**100.00**	**100.00**	**100.00**	**100.00**	76.86	53.13	39.01
Commercial	27.78	27.08	27.78	22.22	83.33	**99.24**	74.24
Road	81.18	80.00	78.82	69.41	81.49	78.10	**84.89**
Parking Lot 1	24.04	7.69	0.00	0.00	0.00	**32.48**	0.00
AA	69.77	68.58	65.63	67.45	78.87	72.45	61.30
OA	64.50	63.22	60.15	61.94	**75.98**	67.45	60.75
κ	0.59	0.58	0.55	0.57	**0.72**	0.63	0.54

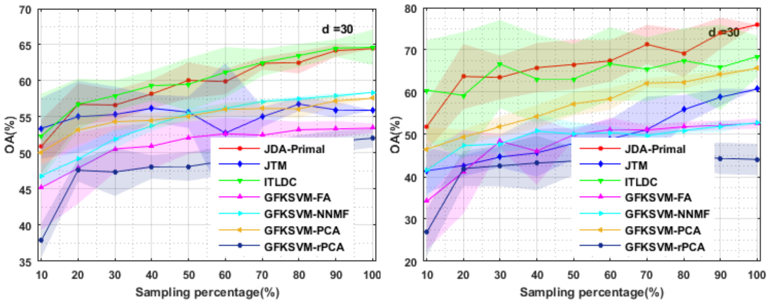
4.4. Influence of the Source Domain Sample Size

Finally, in Figure 16 we present the results of experiments aimed at gauging the influence of the domain samples size used to train the considered DL algorithms. To this end, curves of the overall classification accuracy as a function of the training samples size with a fixed subspace dimensionality (d = 30) and for both the Pavia and the Houston datasets are presented. For more objective evaluation purposes, each experiment was executed 10 times, randomly selecting the training samples from the pool.



(**a**) University→Center (**b**) Center→University

Figure 16. *Cont.*



(**c**) Left→Right (**d**) Right→Left

Figure 16. Classification accuracy curves for GFKSVM and other DA algorithms applied to ROSIS (**a**,**b**) and Houston (**c**,**d**) data. Each curve shows the average overall accuracy value with respect to the increasing size of training sets over 10 independent runs. The shaded areas show the standard deviation of the overall accuracy within the independent runs.

According to the results in Figure 16, one can observe that the proposed GFKSVM with PCA or rPCA shows the best OA values for the ROSIS data, while JDA and ITLDC achieve the best results for the Houston dataset. This indicates again that GFKSVM is more suitable to handle heterogeneous data, while JDA and ITLDC are more suitable for homogeneous data.

5. Conclusions

In order to deal with the data distribution shift problem for the classification of hyperspectral images, in this paper we have experimentally studied the suitability of geodesic flow Gaussian kernel for unsupervised domain adaptation. For comparison purposes, state-of-the-art domain adaptation methods, namely JDA, ITLDC and JTM, were also considered.

Experimental results with two real hyperspectral datasets show that, for relatively statistically "heterogeneous" data sets, the RBF kernel based GFKSVM is more stable with respect to the dimensionality of transferred subspaces, and is capable of learning from small training sets, achieving a better classification performance, especially in the PCA transformed domain. Instead, JDA is more suitable to handle statistically homogeneous datasets, such as the Houston data set, using a large transferred feature dimensionality. Finally, the performances of GFKSVM with rPCA applied to the source and target domains and using a small number of transferred features are not acceptable. Instead, excellent results can be obtained by replacing rPCA with PCA, FA or NMF in one domain.

Future works will be focused on testing the proposed approach in different scenarios, such as domain adaptation between images of different multispectral sensors, as well as in the case of hyperspectral and multispectral sensors as source and target images. Additionally, semi-supervised approaches will be also considered as another interesting line of study.

Acknowledgments: This work was partially supported by 2015 Xinjiang Uygur Autonomous Region High Level Talents Introduction Project (Y648031) and the International Science & Technology Corporation Program of China (010DFA92720, 2012DFF30130). The authors would like to thank editors and anonymous reviewers for their outstanding comments which greatly helped us to improve the quality of our article.

Author Contributions: Alim Samat developed the algorithms, executed all the experiments, finished the original manuscript and following revisions, and provided part of the funding. Paolo Gamba offered valuable suggestions and comments, and revised carefully the original manuscript and the revisions. Jilili Abuduwaili provided part of the funding. Sicong Liu and Zelang Miao, contributed to the revisions of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Camps-Valls, G.; Bruzzone, L. Kernel-based methods for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 1351–1362. [CrossRef]

2. Chi, M.; Bruzzone, L. Semisupervised classification of hyperspectral images by SVMs optimized in the primal. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 1870–1880. [CrossRef]
3. Samat, A.; Du, P.; Liu, S.; Li, J.; Cheng, L. E2LM: Ensemble Extreme Learning Machines for Hyperspectral Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs.* **2014**, *7*, 1060–1069. [CrossRef]
4. Olofsson, P.; Foody, G.M.; Herold, M.; Stehman, S.V.; Woodcock, C.E.; Wulder, M.A. Good practices for estimating area and assessing accuracy of land change. *Remote Sens. Environ.* **2014**, *148*, 42–57. [CrossRef]
5. Bruzzone, L.; Cossu, R. A multiple-cascade-classifier system for a robust and partially unsupervised updating of land-cover maps. *IEEE Trans. Geosci. Remote Sens.* **2002**, *40*, 1984–1996. [CrossRef]
6. Camps-Valls, G.; Gómez-Chova, L.; Muñoz-Marí, J.; Rojo-Álvarez, J.L.; Martínez-Ramón, M. Kernel-based framework for multitemporal and multisource remote sensing data classification and change detection. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 1822–1835. [CrossRef]
7. Bruzzone, L.; Prieto, D.F. Unsupervised retraining of a maximum likelihood classifier for the analysis of multitemporal remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2001**, *39*, 456–460. [CrossRef]
8. Samat, A.; Li, J.; Liu, S.; Du, P.; Miao, Z.; Luo, J. Improved hyperspectral image classification by active learning using pre-designed mixed pixels. *Pattern Recognit.* **2016**, *51*, 43–58. [CrossRef]
9. Izquierdo-Verdiguier, E.; Gómez-Chova, L.; Bruzzone, L.; Camps-Valls, G. Semisupervised kernel feature extraction for remote sensing image analysis. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 5567–5578. [CrossRef]
10. Curlander, J.C. Location of spaceborne SAR imagery. *IEEE Trans. Geosci. Remote Sens.* **1982**, *2*, 359–364. [CrossRef]
11. Sugiyama, M.; Storkey, A.J. Mixture regression for covariate shift. In *Advances in Neural Information Processing Systems*; NIPS proceedings: Vancouver, CO, Canada, 2006; pp. 1337–1344.
12. Huang, J.; Gretton, A.; Borgwardt, K.M.; Schölkopf, B.; Smola, A.J. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems*; NIPS proceedings: Vancouver, CO, Canada, 2006; pp. 601–608.
13. Schott, J.R.; Salvaggio, C.; Volchok, W.J. Radiometric scene normalization using pseudoinvariant features. *Remote Sen. Environ.* **1988**, *26*, 1–16. [CrossRef]
14. Woodcock, C.E.; Macomber, S.A.; Pax-Lenney, M.; Cohen, W.B. Monitoring large areas for forest change using Landsat: Generalization across space, time and Landsat sensors. *Remote Sens. Environ.* **2011**, *78*, 194–203. [CrossRef]
15. Olthof, I.; Butson, C.; Fraser, R. Signature extension through space for northern landcover classification: A comparison of radiometric correction methods. *Remote Sens. Environ.* **2005**, *95*, 290–302. [CrossRef]
16. Rakwatin, P.; Takeuchi, W.; Yasuoka, Y. Stripe noise reduction in MODIS data by combining histogram matching with facet filter. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 1844–1856. [CrossRef]
17. Inamdar, S.; Bovolo, F.; Bruzzone, L.; Chaudhuri, S. Multidimensional probability density function matching for preprocessing of multitemporal remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 1243–1252. [CrossRef]
18. Bruzzone, L.; Marconcini, M. Domain adaptation problems: A DASVM classification technique and a circular validation strategy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 770–787. [CrossRef] [PubMed]
19. Liu, Y.; Li, X. Domain adaptation for land use classification: A spatio-temporal knowledge reusing method. *ISPRS J. Photogramm. Remote Sens.* **2014**, *98*, 133–144. [CrossRef]
20. Banerjee, B.; Bovolo, F.; Bhattacharya, A.; Bruzzone, L.; Chaudhuri, S.; Buddhiraju, K.M. A Novel Graph-Matching-Based Approach for Domain Adaptation in Classification of Remote Sensing Image Pair. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 4045–4062. [CrossRef]
21. Matasci, G.; Volpi, M.; Kanevski, M.; Bruzzone, L.; Tuia, D. Semisupervised Transfer Component Analysis for Domain Adaptation in Remote Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 3550–3564. [CrossRef]
22. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359. [CrossRef]
23. Evgeniou, T.; Pontil, M. Regularized multi-task learning. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2014; pp. 109–117.
24. Daume, H., III; Marcu, D. Domain adaptation for statistical classifiers. *J. Artificial Intell. Res.* **2006**, *26*, 101–126.

25. Foster, G.; Goutte, C.; Kuhn, R. Discriminative instance weighting for domain adaptation in statistical machine translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Cambridge, MA, USA, 9–11 October 2010; pp. 451–459.
26. Jiang, J.; Zhai, C. Instance weighting for domain adaptation in NLP. In Proceedings of the ACL Conference, Prague, Czech Republic, 23–30 June 2007; pp. 264–271.
27. Sugiyama, M.; Nakajima, S.; Kashima, H.; Buenau, P.V.; Kawanabe, M. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems*; NIPS Proceedings: Whistler, BC, Canada, 2007; pp. 1433–1440.
28. Blitzer, J.; McDonald, R.; Pereira, F. Domain adaptation with structural correspondence learning. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Sydney, Australia, 22–23 July 2006; pp. 120–128.
29. Gao, J.; Fan, W.; Jiang, J.; Han, J. Knowledge transfer via multiple model local structure mapping. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, NV, USA, 24–27 August 2008; pp. 283–291.
30. Mihalkova, L.; Huynh, T.; Mooney, R.J. Mapping and revising Markov logic networks for transfer learning. In Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, Vancouver, BC, USA, 22–23 July 2007; Volume 7, pp. 608–614.
31. Nielsen, A.A.; Conradsen, K.; Simpson, J.J. Multivariate alteration detection (MAD) and MAF postprocessing in multispectral, bitemporal image data: New approaches to change detection studies. *Remote Sens. Environ.* **1998**, *64*, 1–19. [CrossRef]
32. Bruzzone, L.; Persello, C. 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.* **2009**, *47*, 3180–3191. [CrossRef]
33. Tuia, D.; Munoz-Mari, J.; Gomez-Chova, L.; Malo, J. Graph matching for adaptation in remote sensing. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 329–341. [CrossRef]
34. Tuia, D.; Volpi, M.; Trolliet, M.; Camps-Valls, G. Semisupervised Manifold Alignment of Multimodal Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 7708–7720. [CrossRef]
35. Pan, S.J.; Tsang, I.W.; Kwok, J.T.; Yang, Q. Domain adaptation via transfer component analysis. *IEEE Trans. Neural Netw.* **2011**, *22*, 199–210. [PubMed]
36. Matasci, G.; Tuia, D.; Kanevski, M. SVM-based boosting of active learning strategies for efficient domain adaptation. *IEEE J. Sel. Top. Appl. Earth Obs.* **2012**, *5*, 1335–1343. [CrossRef]
37. Patel, V.M.; Gopalan, R.; Li, R.; Chellappa, R. Visual Domain Adaptation: A survey of recent advances. *IEEE Signal Process. Mag.* **2015**, *32*, 53–69. [CrossRef]
38. Gong, B.; Shi, Y.; Sha, F.; Grauman, K. Geodesic flow kernel for unsupervised domain adaptation. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 2066–2073.
39. Gopalan, R.; Li, R.; Chellappa, R. Domain adaptation for object recognition: An unsupervised approach. In Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 999–1006.
40. Absil, P.A.; Mahony, R.; Sepulchre, R. Riemannian geometry of Grassmann manifolds with a view on algorithmic computation. *Acta Appl. Math.* **2004**, *80*, 199–220. [CrossRef]
41. Trier, Ø.D.; Jain, A.K.; Taxt, T. Feature extraction methods for character recognition—A survey. *Pattern Recognit.* **1996**, *29*, 641–662. [CrossRef]
42. Lopez-Paz, D.; Sra, S.; Smola, A.; Ghahramani, Z.; Schölkopf, B. Randomized Nonlinear Component Analysis. Available online: http://arxiv.org/abs/1402.0119 (accessed on 9 March 2016).
43. Langville, A.N.; Meyer, C.D.; Albright, R.; Cox, J.; Duling, D. Algorithms, Initializations, and Convergence for the Nonnegative Matrix Factorization. Available online: http://arxiv.org/abs/1407.7299 (accessed on 9 March 2016).
44. Shi, Y.; Sha, F. Information-Theoretical Learning of Discriminative Clusters for Unsupervised Domain Adaptation. Available online: http://arxiv.org/abs/1206.6438 (accessed on 9 March 2016).
45. Long, M.; Wang, J.; Ding, G.; Sun, J.; Yu, P.S. Transfer feature learning with joint distribution adaptation. In Proceedings of the 2013 IEEE International Conference on Computer Vision (ICCV), Sydney, Australia, 1–8 December 2013; pp. 2200–2207.

46. Long, M.; Wang, J.; Ding, G.; Sun, J.; Yu, P.S. Transfer Joint Matching for Unsupervised Domain Adaptation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 1410–1417.
47. Lee, S.I.; Chatalbashev, V.; Vickrey, D.; Koller, D. Learning a meta-level prior for feature relevance from multiple related tasks. In Proceedings of the 24th International Conference on Machine Learning, Corvallis, OR, USA, 20–24 June 2007; pp. 489–496.
48. Pan, S.J.; Kwok, J.T.; Yang, Q. Transfer Learning via Dimensionality Reduction. In Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, Chicago, IL, USA, 13–17 July 2008; Volume 8, pp. 677–682.
49. Hamm, J.; Lee, D.D. Grassmann discriminant analysis: A unifying view on subspace-based learning. In Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 5–9 July 2008; pp. 376–383.
50. Turaga, P.; Veeraraghavan, A.; Srivastava, A.; Chellappa, R. Statistical computations on Grassmann and Stiefel manifolds for image and video-based recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 2273–2286. [CrossRef] [PubMed]
51. Ruiz, A.; López-de-Teruel, P.E. Nonlinear kernel-based statistical pattern analysis. *IEEE Trans. Neural Netw.* **2001**, *12*, 16–32. [CrossRef] [PubMed]
52. Jayasumana, S.; Hartley, R.; Salzmann, M.; Li, H.; Harandi, M. Kernel methods on the riemannian manifold of symmetric positive definite matrices. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013; pp. 73–80.
53. Borgwardt, K.M.; Gretton, A.; Rasch, M.J.; Kriegel, H.P.; Schölkopf, B.; Smola, A.J. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics* **2006**, *22*, 49–57. [CrossRef] [PubMed]