Balanced Weight Joint Geometrical and Statistical Alignment for Unsupervised Domain Adaptation

M. S. Rizal Samsudin, Syed A. R. Abu-Bakar, and Musa M. Mokji
School of Electrical Engineering, Universiti Teknologi Malaysia, Skudai, Johor, Malaysia
Email: ms.rizal1986@graduate.utm.my, {syed, musa}@fke.utm.my

***Abstract*—In real-world applications, images taken from different cameras usually have different resolution, illumination, poses, and background views. This problem leads to the need of domain adaptation in which case, training and testing are not drawn from the same distribution. There have been many studies carried out on domain adaptation, and among the state-of-the-art methods is the Joint Geometrical and Statistical Alignment (JGSA) approach. This paper presents an improvement for unsupervised domain adaptation in transfer learning using a Balanced Weight JGSA (BW-JGSA). The existing method of JGSA seeking the way to minimize the distribution divergence between marginal and conditional distribution across domains; however, treat them equally in terms of distribution weight. This drawback affects the existing method mainly when applied to real applications. The contribution of this paper is to use balanced distribution adaptation in JGSA that can adaptively leverage the importance of marginal and conditional distribution in JGSA. In this method, the balance weight factor, μ, will be applied to marginal and conditional distributions distance for each different subspace in JGSA. Comparing the proposed method with state-of-the-art techniques in object and digital datasets shows significant improvement of our work.**

***Index Terms*—domain adaptation, transfer learning, the balanced weight, joint geometrical, and statistical alignment**

I. Introduction

The development of computer vision and online media applications has attracted many researchers to focus on automatic recognition and analysis of multimedia data due to the success of machine learning techniques to recognize objects or scenes automatically without human assistance. The standard machine learning paradigm is that training and testing data are drawn from the same distribution [1], [2]. However, this assumption does not apply in many applications, especially in the real-world. For instance, in object recognition, the change of background, pixel, angle, and illumination will cause different distribution data between the training and testing phases. This discrepancy will generate a distribution shift, which will affect the accuracy performance because standard classifiers cannot cope with data distribution mismatch. One of the areas that mainly focuses on data distribution mismatch is transfer learning. However, another problem is that labeled data is expensive, and it is unrealistic to relabel a large amount of data in the target domain. Hence, unsupervised domain adaptation, one of the subcategories of transfer learning, is an excellent strategy to leverage the labeled source domain data to boost the new target domain task.

We are focusing on feature transformation in unsupervised learning that transforms the features into joint or unified subspace. Based on the study, there are two ways of feature transformation: data-centric and subspace-centric. The data-centric methods project feature data into a common feature subspace by reducing distribution divergence between the domains. The subspace-centric techniques are based on subspace projection of source and target domain into joint feature subspace. In data-centric methods, two factors influence the better performance; the first is by adapting the marginal and conditional distribution in joint subspace, and the second is leveraging the first concept into two split subspaces by considering the existence of no unified subspace if the dataset has a large discrepancy [3].

Nevertheless, one important factor is to deal with the imbalanced class that often exists in any transfer learning scenarios. The class imbalance depends on the dataset condition; when the dataset is more similar, the conditional distribution is more dominant, and when the dataset is less similar, the marginal distribution is more dominant [4]. Interestingly, the class imbalance problem in unsupervised domain adaptation has been highlighted by [4] through Balanced Distribution Adaptation (BDA). The authors apply manual weight to the marginal and conditional distribution and select the best weight that provides the best accuracy performance. Despite the success, BDA only considers balancing the weight of marginal and conditional distribution in joint subspace that probably gets distorted when there is no unified subspace; a situation when the dataset has a large discrepancy. Furthermore, BDA only focuses on single-centric, one of the data-centric methods, without considering merging with the subspace-centric technique to get better performance.

Manuscript received June 8, 2021; revised December 1, 2021.

doi: 10.12720/jait.13.1.21-28

JGSA in [5] shows that treating the distribution with two disjoint subspaces gives a better result and was apply in many works such as [6]-[8], and this has inspired us to propose the Balanced Weight Joint Geometrical and Statistical Alignment (BW-JGSA). This improvement also applies to the Subspace Alignment (SA), one of the subspace-centric methods. The contributions of this work are two-fold; improving the performance of BDA when dealing with a dataset that has a large discrepancy and boosting the accuracy performance.

Specifically, the idea is to extend the nonparametric Maximum Mean Discrepancy (MMD) to measure the difference in marginal and conditional distributions and integrate with the *balanced weight factor*, μ. Because we consider no unified subspace, the implementation will be different compared to [4], [9] which only produces a single embedded matrix. Instead, there will be two disjointed embedded matrices towards the end. Our improvement will also integrate the subspace alignment to include the subspace-centric to increase the accuracy performance. Based on Fig. 1, we can observe that both domains are projected onto a new feature subspace. This feature subspace is invariant. However, there are risks that the distribution data may not be preserved, different subspace may occur and possibility of class biased due to the geometrical projection. JGSA is known to solve the distribution data and subspace different problem. Our contribution is in handling the class biases by reweighting the data between source domain projection and target domain projection dynamically.

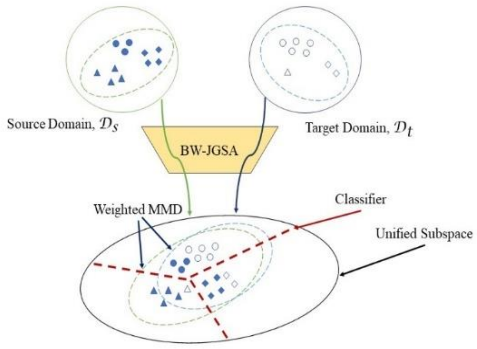


Figure 1. Results of minimizing MMD in BW-JGSA. The class weight biases are handled accordingly, statistical data are preserved and subspace difference are reduced.

The remaining paper is organized as follows: Section II summarizes the existing techniques in unsupervised domain adaptation. Section III details out improved balanced distribution adaptation. Section IV presents the experiment setup, results, and discussion. Finally, the conclusion is discussed in Section V.

II. Related Work

This section discusses prior works on unsupervised domain adaptation related to our work and highlights their differences. We discuss the unsupervised domain adaptation in two groups.

The *Data-centric* group seeks a unified transformation that projects the source and target domains into a new subspace to reduce the discrepancy between the domains and preserve data properties in the original space. Pan *et al.* [10] proposed the Transfer Component Analysis (TCA) that tries to learn some transfer components across domains that minimize the MMD of the new representations of the two domains by using the Reproducing Kernel Hilbert Space (RKHS). Long *et al.* [9] proposed Joint Distribution Adaptation (JDA) that matches marginal and conditional distribution discrepancies between domains. Transfer Joint Matching (TJM) [11] adds instance reweighting into the joint features in TCA. Ghifary *et al.* [12] proposed the Scatter Component Analysis (SCA) that converts feature vectors in the source and target domains into scatter space in the RKHS. Wang *et al.* [4] proposed Balanced Distribution Adaptation (BDA) that leverages the balance weight in the marginal distribution discrepancy and the conditional distribution discrepancy in JDA.

The *Subspace-Centric* group aims to manipulate the domain subspaces in reducing the domain shift without exploiting both the distribution data in the source and target domains. Fernando *et al.* [13] proposed a Subspace Alignment (SA) by aligning the source and target subspaces using a transformation matrix. Gong *et al.* [14] used a manifold property to find a path of the subspace by proposing a Geodesic Flow Kernel (GFK). GFK is performed by embedding the source and target domains using the Grassman manifold and constructing a geodesic manifold between the two points. The infinite number of subspaces were integrated along the flow to construct a path that is later used to form an infinite-dimensional feature vector. The inner product between these feature vectors defines a kernel function that is invariant to the domains. Sun *et al.* [3] proposed a Subspace Distribution Alignment (SDA) to improve both the SA dan GFK by aligning both the source distribution data and the source subpace to the target distribution data in the target subspace. Once this is done, the same classifier used in source domain can be applied directly in the target domain.

Besides the two groups above, Zhang *et al.* [5] proposed to fuse both the data-centric and subspace-centric approaches. In their seminal work, they developed the Joint Geometrical and Statistical Alignment (JGSA) method that splits marginal and conditional distributions into two subspaces and projects these subspaces using the SA technique.

Our work is based on the BDA and the JGSA. However, the BDA is only focusing on improving the class-imbalance in unified subspace and has two drawbacks: (1) it only exploits shared features in two domains, which fails when the two domains have a large discrepancy, and (2) it ignores the importance of the subspace centric. For the JGSA, as mentioned in the introduction, assume the data weight is equally distributed in marginal and conditional distribution. This work will highlight the marginal and conditional distribution imbalance weight problem in JGSA. The

results obtained illustrate the improvement in the JGSA in terms of accuracy in unsupervised domain adaptation.

III. FORMULATION

A. Problem Definition

Definition 1 (Domain). The source domain data denoted as $X_s \in \mathbb{R}^{D \times n_s}$, are drawn from the distribution $P_s(X_s)$ and the target domain denoted as $X_t \in \mathbb{R}^{D \times n_t}$ are drawn from $P_t(X_t)$, where D is the dimension of n_s and n_t, while n_s and n_t are the number of data instances in the source and target domain, respectively. $D_s = \{(x_1, y_1) \ldots (x_{n_s}, y_{n_s})\}$ is defined as the labeled source domain and $D_t = \{(x_1, y_1) \ldots (x_{n_t}, y_{n_t})\}$ is defined as the unlabelled target domain, where $x \in \mathbb{R}^D$.

Definition 2 (Task). The assumption here is that even though the feature space and the label space are the same, i.e. $X_s = X_t$ and $Y_s = Y_t$, due to the dataset shift, their distributions are not. In other words, their marginal distributions would not be the same, i.e. $P_s(X_s) \neq P_t(X_t)$. This is based on a data-centric group that assumes that there is a joint or unified subspace $\emptyset(.)$ that is, however, not valid, particularly when the dataset shift is large. Therefore, we assume $P_s(\emptyset(X_s)) \neq P_t(\emptyset(X_t))$ for the marginal distribution, and $P_s(Y_s|\emptyset(X_s)) \neq P_t(Y_t|\emptyset(X_t))$ for the conditional distribution. The task of the unsupervised domain adaptation is then to learn the labels y_t of D_t by leveraging the source domain D_s.

B. Balanced Weight Joint Geometrical and Statistical Alignment

Since BW-JGSA is based on the conventional JGSA, it basically inherits all of the JGSA properties. Hence, BW-JGSA also (1) minimizes the distribution difference between the two domains and handles the class biases accordingly, (2) reduces the divergence between the source and target subspaces, (3) maximizes the variance of the target domain, and (4) optimizes and preserves the within-class and between-class variances of the source domain. The main difference with the conventional JGSA, however, is the incorporation of the balanced weight factor, μ, in property (1) above.

To reduce the difference between marginal distributions in $P_s(X_s)$ and $P_t(X_t)$, we follow [10] in employing the MMD to compute the distance between the sample mean of the source and target data in the k-dimensional embedding.

$$\min_{A,B} \left\| \frac{1}{n_s} \sum_{x_{si} \in X_s} A^T x_{si} - \frac{1}{n_t} \sum_{x_{tj} \in X_t} B^T x_{tj} \right\|_F^2 \quad (1)$$

To reduce the difference between the conditional distributions $P_s(Y_s|(X_s))$ and $P_t(Y_t|(X_t))$, sufficient labels in the target view are needed. However, since there is no labeled data exists in the target view, we follow Long's [9] technique that utilizes target pseudo labels predicted by the source domain classifier to represent the class-conditional data distributions in the target domain. The pseudo labels in the target domain are iteratively refined to reduce the difference in conditional distributions with the source domains. This method can minimize the conditional distribution shift between domains.

$$\min_{A,B} \sum_{c=1}^{C} \left\| \frac{1}{n_s^{(c)}} \sum_{x_{si} \in X_s^{(c)}} A^T x_{si} - \frac{1}{n_t^{(c)}} \sum_{x_{tj} \in X_t^{(c)}} B^T x_{tj} \right\|_F^2 \quad (2)$$

We then employ balance weight to both the marginal and conditional distributions. In the traditional JGSA, the weight of both the marginal and conditional distributions between domains are equally adjusted. This approach will lead to undesirable bias since marginal distributions are more dominant when they are similar and vice versa. The balance weight exploits a *balance weight factor*, μ to leverage the different contribution of distributions:

$$D(D_s, D_t) \approx (1 - \mu)\, D\left(P_s(X_s), P_t(X_t)\right) + \mu\left(P_s(Y_s|\emptyset(X_s)), P_t(Y_t|\emptyset(X_t))\right) \quad (3)$$

by combining the Eq. (1), (2), and (3), the new representation can be formulated as below:

$$D(D_s, D_t) \approx (1 - \mu) \left\| \frac{1}{n_s} \sum_{x_{si} \in X_s} A^T x_{si} - \frac{1}{n_t} \sum_{x_{tj} \in X_t} B^T x_{tj} \right\|_{\mathcal{H}}^2 + \mu \sum_{c=1}^{C} \left\| \frac{1}{n_s^{(c)}} \sum_{x_{si} \in X_s^{(c)}} A^T x_{si} - \frac{1}{n_t^{(c)}} \sum_{x_{tj} \in X_t^{(c)}} B^T x_{tj} \right\|_{\mathcal{H}}^2 \quad (4)$$

where $\mathcal{H}$ denotes the reproducing kernel Hilbert space (RKHS), $c \in \{1,2,\ldots,C\}$ is the distinct class label. $X_s^{(c)}$ is the set of data instances from class c in the i-th source domain and $n_s^{(c)}$ is the number of data instances $X_s^{(c)}$. Correspondingly, $X_t^{(c)}$ is the set of data instances from class c in the i-th target domain and $n_t^{(c)}$ is the number of data instances $X_s^{(c)}$. The final distribution divergence minimization term can be rewritten as:

$$\min_{A,B} Tr\left(\begin{bmatrix} A^T & B^T \end{bmatrix} \begin{bmatrix} M_{ss} & M_{tt} \\ M_{ts} & M_{st} \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} \right) \quad (5)$$

where

$$M_{ss} = X_s((1 - \mu)L_s + \mu \sum_{c=1}^{C} L_s^{(c)})\, X_s^T$$

$$L_s = \frac{1}{n_s n_s} 1_s 1_s^T,\ (L_s^{(c)})_{ij} = \begin{cases} \frac{1}{n_s^{(c)} n_s^{(c)}} & x_i, x_j \in X_s^{(c)} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

$$M_{tt} = X_t((1 - \mu)L_t + \mu \sum_{c=1}^{C} L_t^{(c)})\, X_t^T$$

$$L_t = \frac{1}{n_t n_t} 1_t 1_t^T, (L_t^{(c)})_{ij} = \begin{cases} \frac{1}{n_t^{(c)} n_t^{(c)}} & x_i, x_j \in X_t^{(c)} \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

$$M_{ts} = X_t((1 - \mu)L_{ts} + \mu \sum_{c=1}^{C} L_{ts}^{(c)})\, X_s^T$$

$$L_{ts} = \frac{-1}{n_t n_s} 1_t 1_s^T,\ (L_{ts}^{(c)})_{ij} = \begin{cases} \frac{-1}{n_t^{(c)} n_s^{(c)}} & x_j \in X_t^{(c)}, x_i \in X_s^{(c)} \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

$$M_{st} = \mathrm{X}_s((1 - \mu)\mathrm{L}_{st} + \mu \sum_{c=1}^{C} L_{st}^{(c)})\, \mathrm{X}_t^T$$

$$\mathrm{L}_{st}=\frac{-1}{n_s n_t} 1_s 1_t^T,\ (L_t^{(c)})_{ij} = \begin{cases} \frac{-1}{n_s^{(c)} n_t^{(c)}} & x_i \in \mathrm{X}_s^{(c)}, x_j \in \mathrm{X}_t^{(c)} \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

For **M**, it can be seen that, instead of measuring the difference and balance the data between cross-domain, M_{ts} and M_{st}, the M_{ss} and M_{tt} allow us to measuring and balance the associated data within source and target domains. To reduce the divergence between source and target subspaces, we use Eq. (10) to shift both the source and target subspaces to be close to one another. Eq. (10) is the geometrical operational to minimize the subspace gap by using the Frobenius norm. In BDA, this term only considers a single embedded matrix Z due to the assumption there is a joint subspace, while in JDA this term of subspace minimizing divergence is not considered.

$$\min_{A,\,B} \|A - B\|_F^2 \quad (10)$$

To maximize the variance of the target domain and preserve the embedded data properties, we use Eq. (11) to achieve this purpose,

$$\max_B Tr(B^T S_t B) \quad (11)$$

where $S = [S_t, \ldots, S_t]$ is obtained by replicating S_t p times. S_t which is defined as $S_t = \mathrm{X}_t \mathrm{H}_t \mathrm{X}_t^T$ is essentially a covariance matrix, $\mathrm{H}_t = \mathrm{I}_t - \frac{1}{n_t} 1_t 1_t^T$ is the centering matrix, while I_t is the identity matrix and $1_t \in \mathbb{R}^{n_t}$ is the column vector with all ones.

To preserve the source domain's discriminative information, we use Eqs. (12) and (13). The discriminative information is important to preserve since the labeled data is only available in the source domain. In this equation, the inter-class is to be maximized, while the intra- class is to be minimized.

$$\max_A Tr(A^T S_b A) \quad (12)$$

$$\min_A Tr(A^T S_w A) \quad (13)$$

$$S_b = \sum_{c=1}^{C} n_s^{(c)} (m_s^{(c)} - \bar{m}_s)(m_s^{(c)} - \bar{m}_s)^T \quad (14)$$

$$S_w = \sum_{c=1}^{C} \mathrm{X}_s^{(c)}\, \mathrm{H}_s^{(c)} (\mathrm{X}_s^{(c)})^T \quad (15)$$

where S_b is inter-class variance matrix and S_w is intra-class variance matrix, both from the source domain data, $\mathrm{X}_s^{(c)} \in \mathbb{R}^{D \times n_s^{(c)}}$ is the set of data instance from class c, $m_s^{(c)} = \frac{1}{n_s^{(c)}} \sum_{i=1}^{n_s^{(c)}} x_i^{(c)}$, $\bar{m}_s = \frac{1}{n_s} \sum_{i=1}^{n_s} x_i$, $\mathrm{H}_s^{(c)} = \mathrm{I}_s^{(c)} - \frac{1}{n_s^{(c)}} 1_s^{(c)} (1_s^{(c)})^T$ is the centering matrix of intra-class data within class c, $\mathrm{I}_s^{(c)} \in \mathbb{R}^{n_s^{(c)} \times n_s^{(c)}}$ is the identity matrix, $1_s \in \mathbb{R}^{n_s^{(c)}}$ is the number of source samples in class c.

Overall, the four criteria above can be formulated by incorporating Eqs. (5), (10), (11), (12), and (13) into a joint objective function and optimization problem as in Eq. (16), as we follow [5]:

$$\max \frac{\alpha\{\text{target var.}\} + \beta\{\text{inter-class var.}\}}{\{\text{distribution shift}\} + \lambda\{\text{subspace shift}\} + \beta\{\text{intra-class var.}\}}$$

$$\max_{A,B} \frac{Tr\left(\begin{bmatrix} A^T & B^T \end{bmatrix} \begin{bmatrix} \beta S_b & 0 \\ 0 & \alpha S_t \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix}\right)}{Tr\left(\begin{bmatrix} A^T & B^T \end{bmatrix} \begin{bmatrix} M_{ss} + \lambda I + \beta S_w & M_{st} - \lambda I \\ M_{ts} - \lambda I & M_{tt} + (\lambda + \alpha) I \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix}\right)} \quad (16)$$

The objective function of Eq. (16) is to encourage the numerator to be maximized and the denominator to be minimized. We also iteratively update the pseudo labels of the target domain data using the learned transformation to improve the labeling quality until convergence.

C. Learning Algorithm

The domain adaptation alignment in geometrical and statistical can be implemented by searching the best projection matrix of A and B. We begin with representing the A and B matrix in Eq. (16) as W. The above optimization can be maximized by treating the denominator to be small to control the solution scale. The above objective function is invariant to rescaling $W \mapsto \alpha W$. Hence, Eq. (16) can be rewritten as

$$\max_W Tr\left(W^T \begin{bmatrix} \beta S_b & 0 \\ 0 & \alpha S_t \end{bmatrix} W\right) \quad (17)$$

$$s.t. Tr\,(W^T \begin{bmatrix} M_{ss} + \lambda I + \beta S_w & M_{st} - \lambda I \\ M_{ts} - \lambda I & M_{tt} + (\lambda + \alpha) I \end{bmatrix} W) = 1$$

According to the constrained optimization theory, we denote $\Phi = diag(\emptyset_1, \ldots\ldots, \emptyset_k) \in \mathbb{R}^{k \times k}$ as the Lagrange multiplier and derive the Lagrange function for Eq. (17) as:

$$\mathrm{L} = Tr\left(W^T \begin{bmatrix} \beta S_b & 0 \\ 0 & \alpha S_t \end{bmatrix} W\right) \quad (18)$$

$$+ Tr((W^T \begin{bmatrix} M_{ss} + \lambda I + \beta S_w & M_{st} - \lambda I \\ M_{ts} - \lambda I & M_{tt} + (\lambda + \alpha) I \end{bmatrix} W - I)\Phi$$

by setting $\frac{\partial L}{\partial W} = 0$, we obtained:

$$\begin{bmatrix} \beta S_b & 0 \\ 0 & \alpha S_t \end{bmatrix} W = \begin{bmatrix} M_{ss} + \lambda I + \beta S_w & M_{st} - \lambda I \\ M_{ts} - \lambda I & M_{tt} + (\lambda + \alpha) I \end{bmatrix} W \Phi \quad (19)$$

where $\Phi = \mathrm{diag}(\emptyset_1, \ldots\ldots, \emptyset_k)$ are the k smallest eigenvectors and $W = [\, W_1, \ldots, W_k \,]$ contains the corresponding eigenvectors. By finding W, the process to obtain the projection matrix of A and B can be solved. Finally, the expected output is the two embedded matrices, $Z_s = A^T \mathrm{X}_s$ and $Z_t = B^T \mathrm{X}_t$.

Kernelization: For nonlinear problems, we follow [9] that applies the kernel mapping $\psi: x \mapsto \psi(x)$, or $\psi(\mathrm{X}) = [\psi(x_1), \ldots, \psi(x_N)]$ and kernel matrix $K = \psi(X)^T \psi(\mathrm{X}) \mathbb{R}^{N \times N}$, where N is the number of all samples in source and target domains. The kernel matrix is constructed using linear or RBF kernel.

IV. Experiments

In this section, we evaluate our proposed approach through a number of experiments.

A. Dataset Preparation

We adopt two widely-used domain adaptation datasets in this paper: (1) The Office+Caltech, and (2) The USPS+MNIST. The Office dataset consists of three real-world object domains: Amazon, Webcam, and DSLR, and contains 1410 images with 10 object categories. Similarly, the Caltech-256 dataset has 10 object categories with 1123 images. The USPS(U) and MNIST(M) are hand-written digits recognition datasets. The USPS(U) dataset consists of 7,291 training images and 2,007 test images of size 16×16 each. The MNIST(M) dataset contains a training set of 60,000 samples and a test set of 10,000 samples of size 28×28 each.

B. Baseline

One non-domain adaptation and seven state-of-the-art domain adaptation approaches were implemented for comparisons purpose, and these are (1) 1-Nearest Neighbour Classifier (1-NN), (2) Subspace Alignment (SA), (3) Subspace Distribution Alignment (SDA), (4) Balanced Distribution Adaptation (BDA), (5) Joint Distribution Adaptation (JDA), (6) Transfer Joint Matching (TJM), (7) Transfer Component Analysis (TCA), and (8) Joint Geometrical and Statistical Alignment (JGSA).

For our work and BDA, balance weight factor μ is searched from {0, 0.1, 0.2, …., 0.9, 1.0}. For the kernel-based methods, we only used linear kernel as implemented in [9]. We followed the common parameter settings as given in [4], i.e. subspace dimension, d =100, regularization parameter, λ = 0.1, and the maximum iteration number, T = 10. In addition, we followed the previous work in [4], [5], [9] such that 1-NN is chosen as the base classifier to produce pseudo-label for the target domain.

We conducted experiments on the Office+Caltech and USPS+MINIST dataset with SURF descriptor with linear kernel for low-level features. Our experiment runs on a PC with an Intel Core i7 CPU (6 cores) and 20 GB RAM. All codings were done using MATLAB.

C. Results and Discussion

- *Classification accuracy*

We begin with the classification accuracy of BW-JGSA along with the seven other methods as shown in Table I and Table II. The results are illustrated in Fig. 2(a) and Fig. 2(b) for better visualization.

For object dataset results given in Table I, we observed that BW-JGSA outperforms all the seven methods on most tasks with 12 out of 12 cross-domain accuracies. On an average-wise, BW-JGSA outperforms others with an average accuracy of 50.61%. We also compared with a non-domain adaptation method, 1-NN, and obtained a 9.69% improvement with BW-JGSA. When compared with BDA and JGSA, the improvement of BW-JGSA is 4.54% and 1.69%, respectively. In general, all the domain adaptation methods outperform the non-domain adaptation method (1-NN). We observed that BW-JGSA outperforms all other methods in 2 out of 2 with an average accuracy of 69.65% for digit datasets.

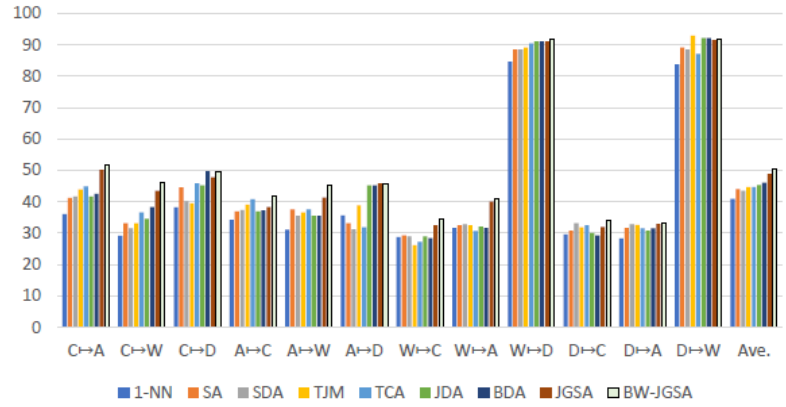
We also observed that TCA performs poorer compared to that of JDA because TCA is not considering the conditional distribution iterations as part of the work. The JDA took both marginal and conditional distribution iterations. However, both performed poorer compared to that of the BDA because of unbalanced weight for marginal and conditional distributions. This is consistent with the previous results obtained in [4]. From the table, it can be seen that the JGSA is the second best because the joint or unified subspace does not exist due to the large discrepancy between domains. However, similar to TCA and JDA, the traditional JGSA does not consider unbalanced weight for marginal and conditional distributions.

TABLE I. ACCURACY (%) BASED ON THE OFFICE+CALTECH256 OBJECT DATASETS

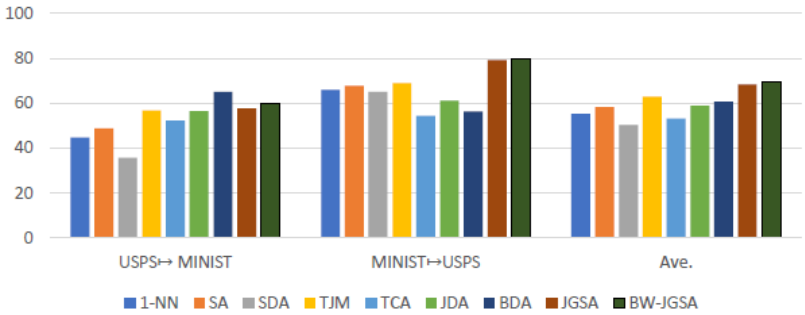
Dataset	1-NN	SA	SDA	TJM	TCA	JDA	BDA	JGSA	BW-JGSA
C↦A	36.01	41.23	41.75	43.95	44.89	41.65	42.48	50.20	**51.88**
C↦W	29.15	33.22	31.53	33.22	36.61	34.58	38.31	43.48	**46.10**
C↦D	38.21	44.59	40.13	39.49	45.86	45.22	49.68	47.77	**49.68**
A↦C	34.19	36.87	37.40	39.09	40.78	36.95	37.31	38.29	**42.03**
A↦W	31.18	37.63	35.59	36.61	37.63	35.59	35.59	41.35	**45.42**
A↦D	35.66	33.12	31.21	38.85	31.85	45.22	45.22	45.85	**45.85**
W↦C	28.76	29.30	28.94	26.18	27.16	28.94	28.41	32.50	**34.55**
W↦A	31.62	32.57	32.88	32.57	30.69	32.05	31.73	40.08	**40.81**
W↦D	84.71	88.54	88.54	89.17	90.45	91.08	91.08	91.08	**91.71**
D↦C	29.56	30.72	33.13	31.88	32.50	30.01	29.30	31.96	**34.11**
D↦A	28.28	31.73	32.78	32.67	31.52	30.79	31.52	32.98	**33.29**
D↦W	83.72	89.15	88.47	92.88	87.12	92.20	92.20	91.52	**91.84**
Average	40.92	44.05	43.53	44.71	44.75	45.36	46.07	48.92	**50.61**

TABLE II. ACCURACY (%) BASED ON THE USPS+MINIST DIGIT DATASETS

Dataset	1-NN	SA	SDA	TJM	TCA	JDA	BDA	JGSA	BW-JGSA
USPS↦MINIST	44.71	48.80	35.70	56.90	52.20	56.50	65.11	57.70	**59.80**
MINIST↦USPS	65.94	67.80	65.00	69.00	54.28	61.22	56.4	79.22	**79.50**
Average	55.32	58.29	50.35	62.95	53.24	58.86	60.76	68.46	**69.65**



(a) Office+Caltech256 object datasets

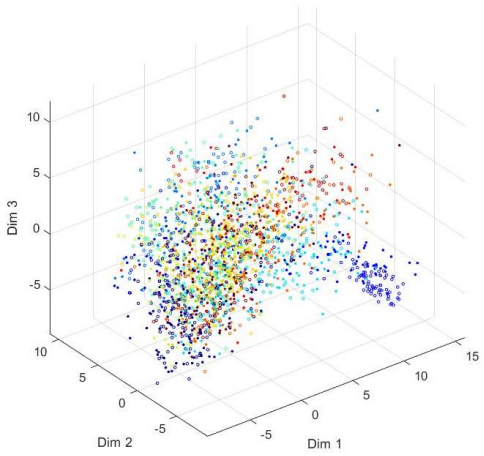


(b) USPS+MINIST digit datasets

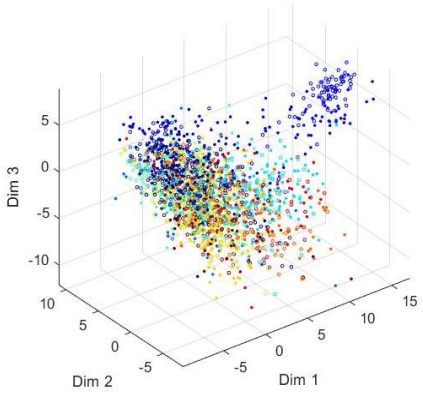
Figure 2. Bar charts of accuracy (%) based on the non-domain adaptation, seven state-of-the-art methods, and our proposed method (BW-JGSA) with different datasets.

- *Visualization of domain adaptation methods*

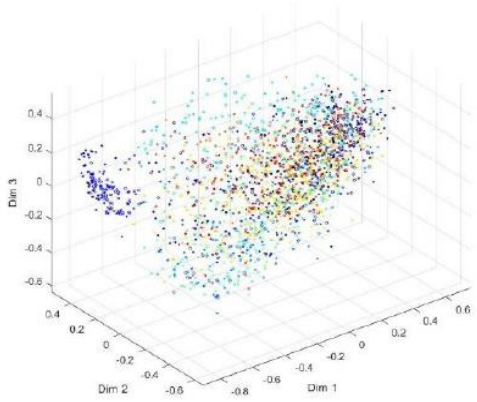
To verify the advantage of BW-JGSA, we plot a selected five domain adaptation methods as well as the BW-JGSA in 3D scatter plots for cross-domain between object datasets Caltech vs. Webcam. From the observation, we can divide the scatter-plots into three categories. The first category is the SA and SDA plots (Fig. 3(a) and Fig. 3(b)) which are based on subspace centric group. We can observe that their distribution difference is still large between domains. This is because the SA and SDA are based on the alignment of subspace and do not preserve the distribution data. The second category is the JDA and BDA plots (Fig. 3(c) and Fig. 3(d)) which are based on data-centric group. JDA considers marginal and conditional distributions to align the data distribution into a joint subspace. BDA is the improvement of JDA that optimizes the balance weight for marginal and conditional distributions. From Fig. 3(c) and Fig. 3(d), we can observe that BDA is smoother in data distribution than JDA. The last category is the JGSA and BW-JGSA plots (Fig. 3(e) and Fig. 3(f)) that consider both subspace-centric and data-centric groups. Although we observe that both data representation is compact, BW-JGSA gives more advantage based on the accuracy performance. JGSA simply optimizes two aligned subspaces, such that the source class information and the target variance can be preserved, and the two subspaces are made to move closer to each other by minimizing the distance between the two. Similar to BDA, BW-JGSA improves the JGSA by optimizing the balance weight, and the visual scatter-3D plot is smoother and balanced.



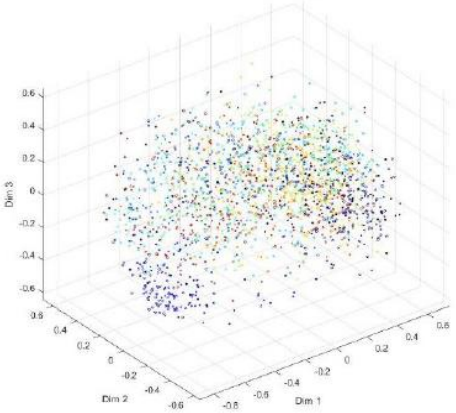
(a) Aligned data classes for SA



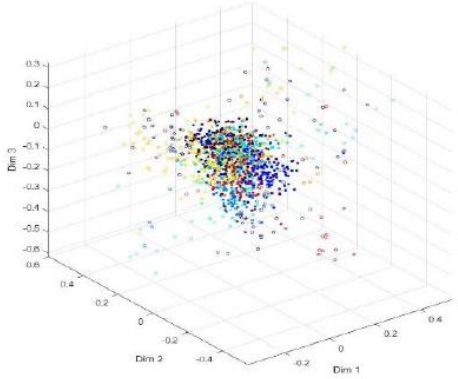
(b) Aligned data classes for SDA



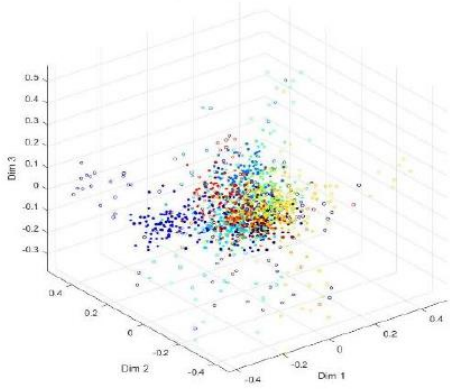
(c) Aligned data classes for JDA



(d) Aligned data classes for BDA



(e) Aligned data classes for JGSA



(f) Aligned data classes for BW-JGSA

Figure 3. 3D scatter plots of source and target for cross-domain object dataset: Caltech vs. Webcam on 10 classes. Dim. 1, 2 and 3 represent the first 3-dimensional feature of the raw SURF features. Filled points represent source domain and empty points represent target domain. (Best view in color).

- *Optimum value of μ*

As mentioned above, the μ value needs to be searched manually from cross-validation with a set of values from {0, 0.1, 0.2, until 1}. The μ value will be different from each cross-domain depending on the dataset that we used. The μ value technically is not a free parameter (i.e., regularization factor, λ), in which μ value has to be estimated according to data distribution.

V. Conclusion

In this paper, we have proposed a Balance Weight-Joint Geometrical and Statistical Alignment (BW-JGSA) that improves the performance of JGSA as one of the state-of-the-art methods in unsupervised domain adaptation. By assuming that marginal and conditional distributions are equally distributed, JGSA fails to capitalize on the importance of each distribution adaptively. The BW-JGSA, on the other hand, gives balance weight to the JGSA by manipulating the balance weight factor, μ in both distributions during the minimization data process. In this work, the balance weight factor is searched manually by maximizing the accuracy performance. The performance has demonstrated that the proposed method improves the JGSA and outperforms the other unsupervised domain adaptation baseline. For future works, we aim to set the value of μ in automation to save the computational time consuming of BW-JGSA.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

All the authors have contributed to this study equally.

Acknowledgment

The authors wish to thank Malaysia Armed Forces and Royal Malaysia Navy for the study sponsor. This work was supported in part by a grant from Ministry of Education Malaysia and University Technology Malaysia (UTM) under the Fundamental Research Scheme, grant number R.J130000.7851.5F179.

References

[1] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345-1359, 2010.

[2] L. Shao, S. Member, F. Zhu, S. Member, and X. Li, "Transfer learning for visual categorization: A survey," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 26, no. 5, pp. 1019-1034, 2015.

[3] B. Sun and K. Saenko, "Subspace distribution alignment for unsupervised domain adaptation," in *Proc. British Machine Vision Conference*, 2015, vol. 4, pp. 24.1-24.10.

[4] J. Wang, Y. Chen, S. Hao, W. Feng, and Z. Shen, "Balanced distribution adaptation for transfer learning," in *Proc. IEEE Int. Conf. Data Mining*, 2017, pp. 1129-1134.

[5] J. Zhang, W. Li, and P. Ogunbona, "Joint geometrical and statistical alignment for visual domain adaptation," in *Proc. 30th IEEE Conf. Comput. Vis. Pattern Recognition*, 2017, pp. 5150-5158.

[6] Y. Liu, Z. Lu, J. Li, and T. Yang, "Hierarchically learned view-invariant representations for cross-view action recognition," *IEEE*

Trans. Circuits Syst. Video Technol., vol. 29, no. 8, pp. 2416-2430, 2019.
[7] J. Zhang, J. Liu, B. Pan, and Z. Shi, "Domain adaptation based on correlation subspace dynamic distribution alignment for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7920-7930, 2020.
[8] R. K. Sanodiya, J. Mathew, R. Aditya, A. Jacob, and B. Nayanar, "Kernelized unified domain adaptation on geometrical manifolds," *Expert Syst. Appl.*, vol. 167, p. 114078, April 2021.
[9] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer feature learning with joint distribution adaptation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2200-2207.
[10] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Networks*, vol. 22, no. 2, pp. 199-210, 2011.
[11] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer joint matching for unsupervised domain adaptation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2014.
[12] M. Ghifary, D. Balduzzi, W. B. Kleijn, and M. Zhang, "Scatter component analysis: A unified framework for domain adaptation and domain generalization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 7, pp. 1414-1430, 2017.
[13] B. Fernando, *et al.*, "Unsupervised visual domain adaptation using subspace alignment," in *Proc. IEEE International Conference on Computer Vision*, 2013, pp. 2960-2967.
[14] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2066-2073.

Mohd Shah Rizal Samsudin received his B.Eng. in Electrical and Electronic Engineering from Universiti Teknologi Malaysia (UTM) in 2009, M.Sc. Electronic System from Universiti Teknologi Petronas (UTP) and M.Sc. Computer Vision from Université de Bourgogne, France in 2015. His research interest includes Digital Image Processing, Machine Learning, Computer Vision and Pattern Recognition. His current research is in the application of transfer learning in human action recognition. Mohd is currently a Ph.D. student, member of the Computer Vision, Video, and Image Processing (CVVIP) research group of the department of Microcomputer and Electronics department at School of Electrical Engineering UTM.

Syed Abdul Rahman Syed Abu Bakar received his B.Sc. (Electrical Engineering) degree from Clarkson University in Potsdam, New York (USA), MSEE degree from Georgia Tech (USA), and the PhD degree from the University of Bradford, England. He has been with the School of Electrical Engineering, Faculty of Engineering, Universiti Teknologi Malaysia since 1992 where he is currently a full professor in the Electronics and Computer Engineering division. His current research interests include Image processing with applications in medical imaging, biometrics, agricultural and industrial applications, watermarking, and computer vision, especially in security and surveillance. He has published more than 150 scientific papers both at national and international levels. He is also a senior member of IEEE.

Musa Mohd Mokji received his B.Eng. in Electrical Engineering from Universiti Teknologi Malaysia. Then he received his M.Eng. and Ph.D. degrees specializing in Image Processing from the same university in 2001 and 2008 respectively. He is currently a senior lecturer at the Faculty of Engineering, UTM with research interest in Signal and Image Processing, Pattern Recognition and Data Mining. Dr. Musa is also interested in the application of these models to agriculture, surveillance system, document processing and medical. He is the head of the Digital Signal and Image Processing research group at the Universiti Teknologi Malaysia. His gives undergraduate and postgraduate lectures on signal processing and image processing at Universiti Teknologi Malaysia.