Preprint Version. Manuscript submitted to and accepted by IEEE Transactions on Cognitive and Developmental Systems

Domain Adaptation Techniques for EEG-based Emotion Recognition: A Comparative Study on Two Public Datasets

Zirui Lan, Olga Sourina, *Senior Member*, *IEEE*, Lipo Wang, *Senior Member*, *IEEE*, Reinhold Scherer, *Member*, *IEEE*, Gernot R. Müller-Putz, *Member*, *IEEE*

***Abstract*—Affective brain-computer interface (aBCI) introduces personal affective factors to human-computer interaction. The state-of-the-art aBCI tailors its classifier to each individual user to achieve accurate emotion classification. A subject-independent classifier that is trained on pooled data from multiple subjects generally leads to inferior accuracy, due to the fact that encephalogram (EEG) patterns vary from subject to subject. Transfer learning or domain adaptation techniques have been leveraged to tackle this problem. Existing studies have reported successful applications of domain adaptation techniques on SEED dataset. However, little is known about the effectiveness of the domain adaptation techniques on other affective datasets or in a cross-dataset application. In this paper, we focus on a comparative study on several state-of-the-art domain adaptation techniques on two datasets: DEAP and SEED. We demonstrate that domain adaptation techniques can improve the classification accuracy on both datasets, but not so effective on DEAP as on SEED. Then, we explore the efficacy of domain adaptation in a cross-dataset setting when the data are collected under different environments using different devices and experimental protocols. Here, we propose to apply domain adaptation to reduce the inter-subject variance as well as technical discrepancies between datasets, and then train a subject-independent classifier on one dataset and test on the other. Experiment results show that using domain adaptation technique in a transductive adaptation setting can improve the accuracy significantly by 7.25% – 13.40% compared to the baseline accuracy where no domain adaptation technique is used.**

***Index Terms*—Electroencephalogram (EEG), affective brain-computer interface (aBCI), domain adaptation, transfer learning, emotion recognition, cross dataset.**

I. Introduction

In recent years, increasing endeavors have been attributed to affective state recognition in the research community of brain-computer interfaces (BCI). An ideal affect-enabled BCI can detect the affective state felt by the user without explicit user input but via spontaneous encephalography (EEG) signals, and respond to different affective states accordingly. Such a BCI could potentially enrich the user experience during the interaction session. Towards that end, various methods have been proposed to identify different affective patterns from brainwaves. The state-of-the-art affective BCIs (aBCI) adopt machine learning techniques and rely on discriminative features [1-2]. A typical aBCI paradigm operates as follows. In a training/calibration session, affective stimuli targeting specific emotions are presented to the user to induce the desired emotions while recording the EEG signals. A classifier is then trained using the chosen features extracted out of the recorded EEG data and the emotion labels. In a live BCI session that immediately follows the training session, the ongoing EEG data are fed to the feature extractor then to the already-trained classifier for real-time emotion classification. Satisfactory classification performance has been reported by many researchers under this paradigm [1]. However, despite encouraging experimental results, the use of an aBCI is still hindered by some factors. Affective EEG patterns vary between different subjects, making it necessary for the subject-of-interest to train a subject-specific classifier. EEG signals are volatile even within the same subject, and a classifier trained at an early time could perform rather poorly at a later time on the same subject. Therefore, frequent recalibrations are needed in order to maintain satisfactory classification accuracy.

In the related field such as motor-imagery BCI, an early attempt to tackle the volatility of the EEG signals was to train the subject to modulate the EEG signals in a way that complies with the classification rule [3-6]. For example, Wolpaw *et. al.* [3] proposed to train the subject to manipulate the mu rhythm power and a movement direction was classified by thresholding the mu power amplitude. The thresholding rule was fixed for the subject, and the subject needed to generate control signals in compliance with the classification rule. They reported high classification accuracy, at the expense of prolonged training time—several weeks. Other attempts involve those adopting transfer learning in a BCI setting [7-10, 13-16]. Transfer learning is a machine learning technique that aims to extract common knowledge from one or more source tasks and apply the knowledge to a related target task [11]. Speaking in a BCI

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its International Research Centers in Singapore Funding Initiative.

Zirui Lan and Olga Sourina are with Fraunhofer Singapore and Nanyang Technological University, Singapore. (e-mail: LanZ0001@e.ntu.edu.sg, EOSourina@ntu.edu.sg).

Lipo Wang is with School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. (Corresponding author, e-mail: ELPWang@ntu.edu.sg).

Reinhold Scherer and Gernot R. Müller-Putz are with Institute of Neural Engineering, Graz University of Technology, Graz, Austria. (e-mail: Reinhold.Scherer@tugraz.at, Gernot.Mueller@tugraz.at).

DOI: 10.1109/TCDS.2018.2826840

context, we can either attempt to find some common feature representations that are invariant across different subjects, or we can try to uncover how the classification rules differ between different subjects. The two methods are denoted as domain adaptation and rule adaptation [12], respectively. Domain adaptation approach has almost exclusively dominated the current BCI-related literature [12]. Krauledat *et. al.* [7] proposed to find prototypical filters of Common Spatial Pattern (CSP) from multiple recording sessions and apply the filters to follow-up session without recalibrating the classifier. Fazli *et. al.* [8] proposed to construct an ensemble of classifiers derived from subject-specific temporal and spatial filters from 45 subjects, and chose a sparse subset of the ensemble that is predictive for a BCI-naïve user. Kang *et. al.* [9] developed composite CSP that is a weighted sum of covariance matrices of multiple subjects to exploit the common knowledge shared between the subjects. Lotte *et. al.* [10] proposed a unifying framework to design regularized CSP that enables subject-to-subject transfer. In aBCI studies, [13-16] explore various domain adaptation methods based on the SEED dataset. In these studies, domain adaptation amounts to finding a domain-invariant space where the inter-subject/inter-session discrepancies of the EEG data are reduced and discriminative features across subjects/sessions are preserved.

Though inter-subject or inter-session transfer and adaptation have been extensively studied in the current literature, the said transfer and adaptation have been restricted within the SEED dataset. That is, the source and target EEG data are from the same dataset in these studies. One question that has not been addressed in the current studies is the efficacy of knowledge transfer and adaptation across different EEG datasets. One could expect that a cross-dataset adaptation sets a more challenging task. Different EEG datasets can be collected using different EEG devices, different experiment protocols, different stimuli etc. These technical differences could add to the discrepancies that are already existing between different subjects/sessions. However, we believe that an ideal, robust BCI should function independently of the device of choice, stimuli used, subjects and experiment context etc. This also makes great practical sense as it relaxes the constraints in a conventional BCI context. Therefore, in this study, we set out to investigate the effectiveness of domain adaptation techniques in a cross-dataset setting, which stands in contrast to existing studies.

Specifically, in this paper, we first investigate the performance of subject-independent emotion recognition with and without domain adaptation techniques in a within-dataset leave-one-subject-out cross-validation setting. We hypothesize that each subject constitutes a domain himself/herself, and that EEG data distribute differently across different domains. We apply several state-of-the-art domain adaptation techniques and compare their performance on DEAP and on SEED datasets. We then propose a cross-dataset emotion recognition scheme to testify the effectiveness of different domain adaptation methods. Under the cross-dataset emotion recognition scheme, the training (source) data are from one dataset and the test (target) data are from the other. Besides the inter-subject variance that is known to exist between different subjects, under a cross-dataset scheme, there also exist technical discrepancies underlying two datasets, hence a more challenging task.

The paper is organized as follows. Section II reviews the two datasets we use in this paper. Section III documents data processing methods, including data preparation, feature extraction, and domain adaptation methods. Section IV explains the experiment in detail. Section V analyzes and discusses the experiment results. The paper is concluded in Section VI.

II. Datasets

There are a few established EEG datasets for affective states investigation. In this paper, we use two of the publicly available datasets, DEAP [19] and SEED [20]. Domain adaptation on SEED has been extensively studied [13-16]. However, little is known about the effectiveness of domain adaptation on DEAP. Moreover, we are also interested in the efficacy of an aBCI in a cross-dataset evaluation setting, especially when two datasets are heterogeneous in many technical aspects. The purpose of cross-dataset evaluation is to attest whether it is possible to maintain satisfactory recognition accuracy when the training data and test data are from different subjects, recorded with different EEG devices, and have the affective states induced by different stimuli, and whether domain adaptation technique can potentially enhance the performance in a cross-dataset evaluation setting.

The DEAP dataset [19] consists of 32 subjects. Each subject was exposed to 40 one-minute long music video as affective stimuli while having the physiological signals recorded. The resultant dataset comprises 32-channel[1] EEG signals, 4-channel Electrooculography (EOG), 4-channel Electromyography (EMG), respiration, plethysmograph, Galvanic Skin Response (GSR) and body temperature. There are 40 EEG trials recorded per subject, each trial corresponding to one emotion elicited by one music video. Immediately after watching each video, the subject was required to rate their truly-felt emotion assessed from five dimensions: valence (associated with pleasantness level), arousal (associated with excitation level), dominance (associated with control power), liking (associated with preference) and familiarity (associated with the knowledge of the stimulus). The rating ranges from one (weakest) to nine (strongest), except familiarity which rates from one to five. The EEG signals were recorded by Biosemi ActiveTwo devices at a sampling rate of 512 Hz and downsampled to 128 Hz.

The SEED dataset [20] contains 15 subjects. Movie excerpts were chosen to elicit three emotions: positive, neutral and negative emotions, with five movie excerpts assigned to each emotion. All subject underwent three EEG recording sessions, with an interval of two weeks between two successive recording sessions. Within each session, each subject was exposed to fifteen four-minute long movie excerpts to induce the desired emotions. The same fifteen movie excerpts were used in all three recording sessions. The resultant dataset contains 15 EEG

[1] The 32 EEG channels include AF3, AF4, C3, C4, CP1, CP2, CP5, CP6, Cz, F3, F4, F7, F8, FC1, FC2, FC5, FC6, Fp1, Fp2, Fz, O1, O2, Oz, P3, P4, P7, P8, PO3, PO4, Pz, T7 and T8.

DOI: 10.1109/TCDS.2018.2826840

trials recorded per subject per session, each emotion having 5 trials. The EEG signals were recorded by a 62-channel[2] ESI NeuroScan device, at a sampling rate of 1000 Hz and downsampled to 200 Hz.

Table I summarizes the technical specifications of the two datasets.

TABLE I TECHNICAL COMPARISONS BETWEEN DEAP AND SEED.

Item	DEAP [19]	SEED [20]
EEG device	Biosemi ActiveTwo	ESI NeuroScan
# of channels	32 for EEG, 8 for peripheral physiological signals	62 for EEG
Sampling rate	Originally 512 Hz, down-sampled to 128 Hz	Originally 1000 Hz, down-sampled to 200 Hz
# of subjects	32	15
Affective stimuli	Music videos	Chinese movie excerpts
Emotions	Valence, liking, arousal, dominance on a 1 (weakest) to 9 (strongest) scale. Familiarity on 1 to 5 scale.	Positive, neutral, negative
# of recording sessions per subject	1	3
# of trials per session	40	15
Trial length	63 seconds	Approx. 4 minutes

III. METHODS

A. Data Preparation

As shown in Table I, the two datasets are quite different in every technical aspect. Given that emotions are rated on five numeric scales in DEAP, we discretize and partition the dimensional emotion space in accordance with SEED as follows: emotions are considered positive if valence rating is greater than 7; neutral if valence rating is smaller than 7 and greater than 3; negative if valence rating is smaller than 3. We then look in DEAP for the trials that have the most participants who reported to have successfully induced positive, neutral and negative emotion, respectively. These trials are: trial #18 for positive emotion, #16 for neutral emotion, and #38 for negative emotion, having 27, 28 and 19 subjects reported that the desired emotion has been induced, respectively. Subjects that commonly reported successful emotion induction with these three trials (#18, #16 and #38) are subjects 2, 5, 10, 11, 12, 13, 14, 15, 19, 22, 24, 26, 28 and 31. Therefore, for DEAP, only the selected trials from these fourteen subjects are used. Each trial lasts for 63 seconds. Since the first 3 seconds are baseline recording without emotion elicitation, we only use the segment from the 4th second to the end. Thus, a valid trial lasts for 60 seconds. For SEED, the trial length varies from 185 seconds to 265 seconds, depending on the length of the affective stimulus used to elicit the desired emotion. We truncate all trials to 185-second-long so as to balance the data of different classes.

B. Feature Extraction

In this study, we adopt differential entropy (DE) as features for emotion recognition. DE features have been extensively used in the current literature [13-16] studying the application of transfer learning techniques in EEG-based emotion recognition. Before feature extraction, each EEG trial is divided into multiple 1-second-long segments. Let T denote one EEG segment, $T \in \mathbb{R}^{s \times w}$, where s is the number of channels, $s = 32$ for DEAP or $s = 62$ for SEED, and w is the number of sampling points per channel per segment, $w = 128$ for DEAP or $w = 200$ for SEED. Each valid trial in DEAP lasts for 60 seconds, and thus yields 60 segments per trial. Similarly, each valid trial in SEED yields 185 segments. The DE feature is extracted out of each EEG segment.

1) Differential Entropy

Let $\boldsymbol{t} \in \mathbb{R}^{w}$ denote the time series of EEG signal from one channel, the DE of $\boldsymbol{t}$ is calculated by [18]

$$\mathrm{DE} = -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(t-\mu)^2}{2\sigma^2}\right) \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(t-\mu)^2}{2\sigma^2}\right) dt$$
$$= \frac{1}{2} \log 2\pi e \sigma^2,$$

where the random variable t follows the Gaussian distribution $N(\mu, \sigma^2)$, and $\boldsymbol{t}$ is the time-series observation of t. The EEG signal, of course, does not follow the Gaussian distribution. It has proven [18] that after $\boldsymbol{t}$ has been band-pass filtered, the time-series of the sub-band signals approximately follow the Gaussian distribution. According to [13-16, 18], five sub-bands are defined: delta (1 – 3 Hz), theta (4 – 7 Hz), alpha (8 – 13 Hz), beta (14 – 30 Hz) and gamma (31 – 50 Hz). As such, five DE features can be extracted from $\boldsymbol{t}$. The final feature vector is a concatenation of features from all channels. For DEAP, the final feature vector is of $5 \times 32 = 160$ dimensions, and each trial yields 60 samples. For SEED, the final feature vector is of $5 \times 62 = 310$ dimensions, and each trial yields 185 samples.

C. Domain Adaptation Methods

In the following, we assume that we have a set of labeled data $X_s \in \mathbb{R}^{m \times n_s}$ and a set of unlabeled data $X_t \in \mathbb{R}^{m \times n_t}$, where m is the dimension of the feature, n_s and n_t are the number of samples in the respective set. Let Y_s be the labels associated with X_s, we refer to $\mathcal{D}_s = \{(X_s, Y_s)\}$ as the source domain, and $\mathcal{D}_t = \{X_t\}$ the target domain. In many use cases, X_s and X_t are differently distributed. That said, domain discrepancies exist between the source and the target domain. Usually, a classifier trained in $\mathcal{D}_s$ can perform rather poorly when directly applied to $\mathcal{D}_t$. The task of domain adaptation is to find a latent, domain-invariant subspace to project $X = [X_s\, X_t] \in \mathbb{R}^{m \times n}$ to be $X' = [X'_s\, X'_t] \in \mathbb{R}^{h \times n}$, where h is the desired dimension of the latent subspace, and $n = n_s + n_t$. In the domain invariant subspace, the discrepancies between X'_s and X'_t have been reduced. Subsequently, we can train a classifier in $\mathcal{D}'_s = \{(X'_s, Y_s)\}$ and apply it to $\mathcal{D}'_t = \{X'_t\}$. This is a typical unsupervised transductive transfer learning setting [11].

[2] The 62 EEG channels include AF3, AF4, C1, C2, C3, C4, C5, C6, CB1, CB2, CP1, CP2, CP3, CP4, CP5, CP6, CPZ, CZ, F1, F2, F3, F4, F5, F6, F7, F8, FC1, FC2, FC3, FC4, FC5, FC6, FCZ, FP1, FP2, FPZ, FT7, FT8, FZ, O1, O2, OZ, P1, P2, P3, P4, P5, P6, P7, P8, PO3, PO4, PO5, PO6, PO7, PO8, POZ, PZ, T7, T8, TP7 and TP8.

1) Maximum Independence Domain Adaptation

Maximum Independence Domain Adaptation (MIDA) [17] seeks to maximize the independence between the projected samples and their respective domain features measured by the Hilbert-Schmidt Independence Criterion (HSIC) [24]. Domain feature captures the background information of a specific sample, for example, which domain the sample belongs to. The domain feature $\boldsymbol{d} \in \mathbb{R}^{m_d}$ of a specific sample $\boldsymbol{x} \in \mathbb{R}^m$ is defined using one-hot encoding scheme as $d_i = 1$ if the sample is from subject i, and 0 otherwise, where m_d is the number of subjects considered, d_i the ith element of $\boldsymbol{d}$. In a cross-dataset scheme, $d_i = 1$ if subject i is from DEAP, or $d_{i+14} = 1$ if subject i is from SEED, and 0 otherwise. The first fourteen bits of $\boldsymbol{d}$ are attributed to subjects from DEAP dataset, and the remaining fifteen bits attributed to subjects from SEED dataset. The feature vector is augmented with its domain feature by concatenation $\hat{\boldsymbol{x}} = [\boldsymbol{x}^\top\ \boldsymbol{d}^\top]^\top \in \mathbb{R}^{m+m_d}$. By augmenting the feature vector with domain feature, we need not distinguish which domain a specific sample is from, and such information is encoded in the augmented feature vector.

Let $\hat{X} = \begin{bmatrix} X \\ D \end{bmatrix} \in \mathbb{R}^{(m+m_d)\times n}$ be the matrix of the augmented feature where source data and target data are pooled together, we project $\hat{X}$ to the desired subspace by applying a mapping ϕ followed by a linear transformation matrix $\widetilde{W}$ to $\hat{X}$, denoted by $X' = \widetilde{W}^\top \phi(\hat{X})$. Like other kernel dimensionality reduction methods [26-27], the key idea is to construct $\widetilde{W}$ as a linear combination of all samples in $\phi(\hat{X})$, namely $\widetilde{W} = \phi(\hat{X})W$. Hence, $X' = W^\top \phi(\hat{X})^\top \phi(\hat{X})$. Using the kernel trick, we need not compute $\phi(\hat{X})^\top \phi(\hat{X})$ explicitly in the ϕ space, but in the original feature space via a proper kernel function $\ker(\cdot)$. Let $K_{\hat{X}} = \phi(\hat{X})^\top \phi(\hat{X}) \in \mathbb{R}^{n\times n}$ denote the kernel matrix of $\hat{X}$, $K_{\hat{X}} = [k_{ij}]$, where k_{ij} is computed by $k_{ij} = \ker(\hat{X}_{:i}, \hat{X}_{:j})$, where $\hat{X}_{:i}$ is the ith column of $\hat{X}$, and $\ker(\boldsymbol{u},\boldsymbol{v})$ is a proper kernel function that can take the form of linear function ($\ker(\boldsymbol{u},\boldsymbol{v}) = \boldsymbol{u}^\top\boldsymbol{v}$), polynomial function ($\ker(\boldsymbol{u},\boldsymbol{v}) = (\boldsymbol{u}^\top\boldsymbol{v} + c)^d$), or radial basis function (RBF, $\ker(\boldsymbol{u},\boldsymbol{v}) = \exp(-\frac{\|\boldsymbol{u}-\boldsymbol{v}\|}{2\sigma^2})$) etc.

$W \in \mathbb{R}^{n\times h}$ is the actual projection matrix we wish to find, and such matrix should bear the desired property so that after projection, X' is independent of domain feature D. Intuitively, when X' is independent of domain features D, we cannot distinguish from which domain a specific sample $X'_{:i}$ comes, suggesting that the difference of distribution among different domains is reduced in X'. The HSIC [24] is used as a convenient method to quantify the level of independence. $\mathrm{HSIC}(X', D) = 0$ if and only if X' and D are independent [28]. The larger the HSIC value is, the stronger dependence. HSIC has a convenient but biased empirical estimate given by $(n-1)^{-2}\mathrm{tr}(K_{X'} H K_D H)$ [24], where $K_D = D^\top D \in \mathbb{R}^{n\times n}$ and $K_{X'} = (W^\top K_{\hat{X}})^\top (W^\top K_{\hat{X}}) \in \mathbb{R}^{n\times n}$ are the kernel matrices of X' and D, respectively, $H = I - n^{-1}\mathbf{1}_n\mathbf{1}_n^\top \in \mathbb{R}^{n\times n}$ is the centering matrix, and $\mathbf{1}_n$ is an all-one vector of dimension n.

Besides maximizing the independence between the projected samples and the domain features, it is also important to preserve the statistical property of the data in the latent space, such as the variance [25]. This can be done by maximizing the trace of the covariance matrix $\mathrm{cov}(X') = \frac{1}{n}(X' - \overline{X'})(X' - \overline{X'})^\top$ of the projected samples, where $\overline{X'}$ denotes the mean of X'. Assembling the HSIC (dropping the scalar) and the covariance objectives, and further adding an orthogonal constraint on W, the final objective function to be maximized is

$$\begin{aligned} &\max_W -\mathrm{tr}(W^\top K_{\hat{X}} H K_D H K_{\hat{X}} W) + \mu\mathrm{tr}(W^\top K_{\hat{X}} H K_{\hat{X}} W), \\ &\text{s.t.}\ \ W^\top W = I, \end{aligned} \tag{1}$$

where $\mu > 0$ is a trade-off parameter between optimizing the HSIC and the covariance. The solution of W is given by the h eigenvectors of $K_{\hat{X}}(-HK_DH + \mu H)K_{\hat{X}}$ corresponding to the h largest eigenvalues.

2) Transfer Component Analysis

Transfer Component Analysis (TCA) [29] attempts to mitigate the distribution mismatch by minimizing the Maximum Mean Discrepancy (MMD) in a reproducing kernel Hilbert space (RKHS) [34], which measures the distance between the empirical means of the source domain and the target domain. Intuitively, when the distance between the means of both domains is small, the data tend to distribute similarly in both domains. It has proven that when the RKHS is universal, MMD will asymptotically approach zero if and only if the two distributions are identical [35]. Using the kernel trick, the distance measured in terms of MMD between the means of the projected source data X'_s and target data X'_t in the latent subspace evaluates to

$$\mathrm{Dist}(X'_s, X'_t) = \mathrm{tr}\big((KWW^\top K)L\big) = \mathrm{tr}(W^\top KLKW), \tag{2}$$

where $W \in \mathbb{R}^{n\times h}$ is the projection matrix, $K = [k_{ij}] \in \mathbb{R}^{n\times n}$ the kernel matrix defined on X, and $L = [L_{ij}]$ where $L_{ij} = 1/n_s^2$ if $X_{:i}, X_{:j} \in X_s$, else $L_{ij} = 1/n_t^2$ if $X_{:i}, X_{:j} \in X_t$, otherwise $L_{ij} = -(1/n_s n_t)$.

The cost function comprises the distance and a regularization term we wish to minimize, and is subjected to a variance constraint [29]:

$$\begin{aligned} &\min_W \mathrm{tr}(W^\top KLKW) + \mu\mathrm{tr}(W^\top W), \\ &\text{s.t.}\ \ W^\top KHKW = I. \end{aligned} \tag{3}$$

where $H \in \mathbb{R}^{n\times n}$ is the same centering matrix as in MIDA, and μ the trade-off parameter. Solving (3) for W analytically yields the h eigenvectors of $(KLK + \mu I)^{-1}KHK$ corresponding to the h leading eigenvalues.

3) Subspace Alignment

Subspace alignment (SA) [30] attempts to align the principal component analysis (PCA)-induced bases of the subspace of the source and the target domains. We generate the bases of the h-dimensional subspaces of the source domain and the target domain by applying PCA to X_s and X_t and taking the h eigenvectors corresponding to the h leading eigenvalues. Let Z_s and Z_t denote the bases of the subspaces of the source and the target domain, respectively, $Z_s \in \mathbb{R}^{m\times h} = \mathrm{PCA}(X_s, h)$, $Z_t \in \mathbb{R}^{m\times h} = \mathrm{PCA}(X_t, h)$. To align Z_s with Z_t, a linear transformation matrix $W \in \mathbb{R}^{h\times h}$ is applied to Z_s. The desired W is to minimize the Bregman matrix divergence:

$$W = \min_W (\|Z_s W - Z_t\|_{\mathcal{F}}^2), \tag{4}$$

where $\|\cdot\|_{\mathcal{F}}^2$ is the Frobenius norm. It follows that the closed-form solution of W is given by $W = Z_s^\top Z_t$ [30]. The source and target data can then be projected to the aligned subspaces, respectively, by $X_s' = X_s Z_s Z_s^\top Z_t$ and $X_t' = X_t Z_t$.

4) *Information Theoretical Learning*

Information theoretical learning (ITL) [31] hypothesizes discriminative clustering and consists in optimizing two information-theoretical quantities: (6) and (7).

Let $W \in \mathbb{R}^{h\times m}$ be the projection matrix to the domain-invariant subspace. The squared distance between two points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in the subspaces is expressed as $d_{ij}^2 = \|W\boldsymbol{x}_i - W\boldsymbol{x}_j\|_2^2 = (\boldsymbol{x}_i - \boldsymbol{x}_j)^\top M(\boldsymbol{x}_i - \boldsymbol{x}_j)$, where $M = W^\top W$ is the Mahalanobis distance metric in the original m-dimensional feature space. Given a point $\boldsymbol{x}_i$ and a set of points $\{\boldsymbol{x}_j\}$, the conditional probability of having $\boldsymbol{x}_j$ as the nearest neighbor of $\boldsymbol{x}_i$ is parametrized by $p_{ij} = e^{-d_{ij}^2} / \sum_{j\neq i} e^{-d_{ij}^2}$. Thus, if the labels of $\{\boldsymbol{x}_j\}$ are known (e.g., $\{\boldsymbol{x}_j\}$ are from the source data), it follows that the posterior probability $\hat{p}(y_i = k|\boldsymbol{x}_i)$ for labeling $\boldsymbol{x}_i$ as class k is

$$\hat{p}_{ik} = \textstyle\sum_{j\neq i} p_{ij}\delta_{jk}, \tag{5}$$

where δ_{jk} is 1 if $\boldsymbol{x}_j$ is labeled as class k, and 0 otherwise. Given c classes, a c-dimensional probability vector can be formed: $\hat{\boldsymbol{p}}_i = [\hat{p}_{i1}, \hat{p}_{i2}, \dots, \hat{p}_{ic}]^\top$. We wish to maximize the mutual information between the target data X_t and their estimated labels $\hat{Y}$ parametrized by $\hat{\boldsymbol{p}}$:

$$I_t(X_t; \hat{Y}) = H[\hat{\boldsymbol{p}}_0] - \tfrac{1}{n_t}\textstyle\sum_i H[\hat{\boldsymbol{p}}_i], \tag{6}$$

where $H[\boldsymbol{p}] = -\sum_i p_i \, log \, p_i$ denotes the entropy of the probability vector $\boldsymbol{p}$, $\hat{\boldsymbol{p}}_0$ is the prior distribution given by $1/n_t \sum_i \hat{\boldsymbol{p}}_i$.

Since $\hat{\boldsymbol{p}}_i$ is estimated based on the principle of nearest neighbors, the validity of $\hat{\boldsymbol{p}}_i$ hinges on the assumption that the source data and target data are close to each other in the latent subspace. That said, given a sample $\boldsymbol{x}_i$ and a binary probability vector $\boldsymbol{q}_i$ denoting its domain label, if the assumption holds, we cannot determine $\boldsymbol{q}_i$ given $\boldsymbol{x}_i$ well above the chance level. To achieve this, we minimize the mutual information between domain label Q and data samples X, expressed as

$$I_{st}(X; Q) = H[\hat{\boldsymbol{q}}_0] - \tfrac{1}{n}\textstyle\sum_i H[\hat{\boldsymbol{q}}_i], \tag{7}$$

where $\hat{\boldsymbol{q}}_i = [\hat{q}_{i1} \; \hat{q}_{i2}]^\top$ is estimated via $\hat{q}_{ik} = \sum_{j\neq i} p_{ij}\delta_{jk}$ similar to (5), except that δ_{jk} now indicates domain label. The prior probability $\hat{\boldsymbol{q}}_0$ is computed as $\frac{1}{n}\sum_i \hat{\boldsymbol{q}}_i$.

Assembling the two information-theoretical quantities, we derive the cost function as

$$\min_W -I_t(X_t; \hat{Y}) + \lambda I_{st}(X; Q), \tag{8}$$

where λ is a trade-off parameter. The cost function (8) is parametrized by W and is a non-convex function. We resort to iterative gradient descend methods to optimize (8). W can be heuristically initialized as being the PCA of the target domain [31].

5) *Geodesic Flow Kernel*

The subspaces of the source and the target domains are represented by two points on a Grassmann manifold, where geometric, differential, and probabilistic structures can be defined [32]. Authors of Geodesic Flow Kernel (GFK) domain adaptation [32] proposed to construct a geodesic flow linking the subspaces of the source and the target domain via an infinite number of interpolating subspaces in-between on a Grassmann manifold. Then, they project the source and the target data into each of the infinitely many interpolating subspaces and concatenate the resultant, infinitely many feature vectors to form a super feature vector. To avoid explicitly manipulating on this infinite dimensional feature space, they leverage geodesic flow kernel representing the inner products between any two points in the infinite space, known as the kernel trick. Let $\boldsymbol{x}_i$, $\boldsymbol{x}_j$ be two points in the original m-dimensional feature space, the GFK between them is defined as

$$\mathrm{GFK}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^\top G \boldsymbol{x}_j, \tag{9}$$

To derive G, we need some more math definitions. Let P_s, $P_t \in \mathbb{R}^{m\times h}$ be the bases of the subspaces induced by PCA for the source data and the target data, $R_s \in \mathbb{R}^{m\times(m-h)}$ be the orthogonal complement to P_s, namely $R_s^\top P_s = 0$. Let $U_1 \in \mathbb{R}^{h\times h}$, $U_2 \in \mathbb{R}^{(m-h)\times h}$ be the components of the following pair of singular value decomposition (SVD),

$$P_s^\top P_t = U_1 \Gamma V^\top, \; R_s^\top P_t = -U_2 \Sigma V^\top. \tag{10}$$

Γ and Σ are $h \times h$ diagonal matrices consisting of $\cos\theta_i$ and $\sin\theta_i$ for $i = 1, 2, \dots, h$, where θ_i are the principal angles between the ith bases of P_s and P_t. Then, G is defined as

$$G = [P_s U_1 \; R_s U_2] \begin{bmatrix} \Lambda_1 & \Lambda_2 \\ \Lambda_2 & \Lambda_3 \end{bmatrix} \begin{bmatrix} U_1^\top P_s^\top \\ U_2^\top R_s^\top \end{bmatrix}, \tag{11}$$

where Λ_1, Λ_2 and Λ_3 are diagonal matrices consisting of diagonal elements

$$\lambda_{1i} = 1 + \frac{\sin(2\theta_i)}{2\theta_i}, \lambda_{2i} = \frac{\cos(2\theta_i)-1}{2\theta_i}, \lambda_{3i} = 1 - \frac{\sin(2\theta_i)}{2\theta_i}. \tag{12}$$

6) *Kernel Principal Component Analysis*

Kernel-based principal component analysis (KPCA) [33] is the kernelized extension of PCA exploiting the kernel trick. Strictly speaking, KPCA was not originally developed for domain adaptation purpose, but has been included for comparison with domain adaptation methods in the literature [13-14, 17, 29], citing the denoising and dimension reduction effect of KPCA. Let ϕ be the mapping that maps X to a possibly very high dimensional space, the kernel matrix $K_X = [k_{ij}] = \phi(X)^\top \phi(X)$ for X is computed with a proper kernel function $\ker(\cdot)$, $k_{ij} = \ker(X_{:i}, X_{:j})$. We then center the kernel matrix K by computing $\widetilde{K} = K - HK - KH + HKH$, where $H \in \mathbb{R}^{n\times n}$ is the centering matrix with all elements equal to $1/n$. Next, we solve the eigendecomposition problem $\lambda V = \widetilde{K} V$ for V and λ, where V and λ denote the eigenvectors and eigenvalues, respectively.

Let h be the desired latent subspace dimension, the projection matrix W is constructed with the h eigenvectors from V corresponding to the h largest eigenvalues. The projected samples $X' = [X_s' \; X_t']$ are computed by $X' = K_X W$. It has proven

[33] that KPCA is equivalent to performing standard PCA in the ϕ space, directly manipulating which can be prohibitively expensive.

Before we proceed to the next section, we briefly discuss the distinctions of the methods. MIDA is the only method that can handle multiple source domains, thanks to its domain feature augmentation. MIDA and TCA are closely related in that both try to optimize statistics in RKHS. MIDA, TCA, GFK, and KPCA employ kernel methods and transform the data into kernel representation. GFK and SA have closed-form solutions, which gives them advantages in speed. ITL is based on iterative gradient optimization and may be slower than other methods. It is worth pointing out that the label information Y_s is not used in any method, and the transfer learning is carried out on an unsupervised basis.

IV. Experiments

In the following experiments, we first evaluate the effectiveness of the domain adaptation techniques in a within-dataset leave-one-subject-out cross-validation setting. We then focus on the evaluation of cross-dataset domain adaptation performance. Both evaluations are based on an unsupervised transductive transfer learning scheme [11] as was used in [13-16]. In both settings, we apply logistic regression classifier for classification.

A. Within-dataset domain adaptation

In this experiment, we evaluate the classification accuracy on a leave-one-subject-out cross-validation basis. Specifically, one subject from the dataset in question is left out as the test subject, and the remaining subjects are viewed as training subjects who contribute training data. We hypothesize that each subject constitutes his own domain, thus we have multiple source domains. MIDA is capable of handling multiple source domains. For other methods, we pool together all source domains to form a super source domain. In DEAP, one subject contributes 180 samples (60 samples/class). As such, the training set consists of $180 \times 13 = 2340$ samples from 13 subjects, and the test set 180 samples from the test subject. In SEED, one subject contributes 2775 samples (925 samples/class/session). The training set comprises $2775 \times 14 = 38850$ samples from 14 subjects and the test set 2275 samples from the test subject. We adopt unsupervised transductive domain adaptation scheme to jointly project the training data and test data to the latent, domain-invariant subspace. It has to be pointed out that for SEED, due to the large number of training samples, it is infeasible to include all training samples into the domain adaptation algorithm given the limited computer memory [13-14]. Therefore, for SEED, we randomly sample 1/10 of the training data, equaling to 3885 samples, as actual training data for the domain adaptation algorithms and the subsequent classifier training. We repeat the procedure 10 times for SEED, so that the randomly sampled training data covers a good range of the whole training data set. The classification performance is averaged over 10 runs. We compare the performance of several state-of-the-art domain adaptation techniques to each other, as well as to the baseline performance where no domain adaptation method is adopted. We also compare the domain adaptation performance on two established affective EEG datasets. We stress that domain adaptation techniques have been applied to SEED with success in [13-16]. However, there is little study looking into the performance of domain adaptation techniques on DEAP. Chai *et al.* [16] mentioned briefly without presenting results that it is difficult to successfully apply domain adaptation techniques on DEAP, and that negative transfer has been observed, where the classification performance is actually degraded when domain adaptation techniques are applied.

Table II. Details of hyperparameters

Method	Hyperparameters
MIDA [17]	Kernel = linear, $\mu = 1$, $h = \{5, 10, \dots, 100\}$
TCA [29]	Kernel = linear, $\mu = 1$, $h = \{5, 10, \dots 100\}$
SA [30]	$h = \{5, 10, \dots, 100\}$
ITL [31]	$\lambda = 1$, $h = \{5, 10, \dots, 100\}$
GFK [32]	$h = \{5, 10, \dots, 100\}$
KPCA [33]	Kernel = linear, $h = \{5, 10, \dots, 100\}$

As with other machine learning algorithms, domain adaptation algorithms require that certain hyperparameters be set. One such common hyperparameter is the dimension of the latent subspace. We find the best latent dimension h by searching $\{5, 10, \dots, 100\}$ for each domain adaptation algorithm, respectively. For other hyperparameters, we set to the default values recommended by their authors. Table II gives the details of the hyperparameters used in this experiment.

Table III presents the classification accuracy of different methods on DEAP and SEED. For DEAP, the mean classification accuracy (std) of the baseline method is 39.05% (8.36). Note that the theoretical chance level for random guessing is 33.33 %, and the baseline accuracy is seemingly close to random guess. The real chance level is dependent on the classifier and the number of test samples [36][37]. When there are infinitely many samples, the real chance level approaches the theoretical value. For a finite number of samples, the chance level is computed based on repeated simulations of classifying samples with randomized class labels, as is suggested in [36][37]. We carry out the chance level simulation and present also in Table III the upper bound of the 95 % confidence interval of the accuracy of simulated random guessing. As we can see, the baseline accuracy exceeds the upper bound of the chance level, which leads to the assertion that the baseline is significantly better than chance at a 5 % significance level. Nonetheless, the low absolute accuracy still suggests that there are substantial discrepancies between the sample distributions of different subjects, without handling which would adversely affect the classification accuracy. SA yields an accuracy slightly inferior to the baseline (38.73% vs. 39.05%), suggesting that negative transfer may have happened. Other domain adaptation methods yield improved classification performance over baseline performance. MIDA sees a 9.88 % improvement over the baseline and is the best-performing method, closely followed by TCA. Though the relative improvement is significant (t-test, $p < 0.05$), the absolute accuracy is still rather low. On SEED, the baseline accuracies are noticeably higher than that on DEAP, and much higher than

TABLE III. WITHIN-DATASET LEAVE-ONE-SUBJECT-OUT CROSS-VALIDATION ACCURACY, MEAN % (STD %).

Method	DEAP	SEED			
		Session I	Session II	Session III	Session Average
Baseline	39.05 (8.36)	57.96 (10.85)	48.79 (14.47)	57.45 (13.09)	54.73 (12.80)
MIDA [17]	**48.93** (15.5)	72.31 (10.86)	**69.45** (15.18)	**75.64** (11.37)	**72.47** (12.47)
TCA [29]	47.22 (15.59)	**73.56** (8.37)	68.89 (14.43)	72.57 (11.38)	71.67 (11.3)
SA [30]	38.73 (9.39)	66.03 (7.49)	64.14 (10.47)	66.67 (10.59)	65.61 (9.52)
ITL [31]	40.56 (11.92)	65.82 (11.86)	64.00 (15.09)	69.08 (14.77)	66.30 (13.91)
GFK [32]	46.51 (13.48)	65.75 (12.06)	64.15 (12.12)	72.62 (12.87)	67.51 (12.35)
KPCA [33]	39.84 (11.37)	63.56 (11.01)	58.34 (11.51)	65.58 (9.80)	62.49 (10.77)
Acc Diff (Best)	9.88	15.6	20.66	18.19	17.74
Upp Bnd of Chn Lvl	38.85	34.58	34.65	34.60	34.61

TABLE IV. CROSS-DATASET LEAVE-ONE-SUBJECT-OUT CROSS-VALIDATION ACCURACY, MEAN % (STD %).

Method	DEAP→SEED I	DEAP→SEED II	DEAP→SEED III	SEED I→DEAP	SEED II→DEAP	SEED III→DEAP
Baseline	34.42 (2.82)	34.71 (3.91)	33.71 (3.94)	34.57 (7.98)	32.99 (3.44)	32.51 (6.73)
MIDA [17]	43.75 (10.04)	46.68 (8.01)	**47.11** (10.60)	40.34 (14.72)	39.90 (14.83)	37.46 (13.11)
TCA [29]	**46.95** (11.77)	**47.68** (10.12)	45.83 (11.01)	**42.60** (14.69)	**42.40** (14.56)	**39.76** (15.15)
SA [30]	46.74 (9.47)	42.35 (8.71)	40.65 (12.37)	36.73 (10.69)	37.36 (7.90)	37.27 (10.05)
ITL [31]	42.69 (8.77)	45.19 (7.05)	44.94 (9.05)	34.50 (13.17)	34.10 (9.29)	33.62 (10.53)
GFK [32]	39.53 (4.57)	38.98 (4.76)	39.67 (6.43)	41.91 (11.33)	40.08 (11.53)	39.53 (11.31)
KPCA [33]	42.98 (12.41)	44.26 (11.49)	39.99 (11.29)	35.60 (6.97)	34.69 (4.34)	35.11 (10.05)
Acc Diff (Best)	12.53	12.97	13.40	8.03	9.41	7.25
Upp Bnd of Chn Lvl.	34.68	34.72	34.74	38.35	38.38	38.44

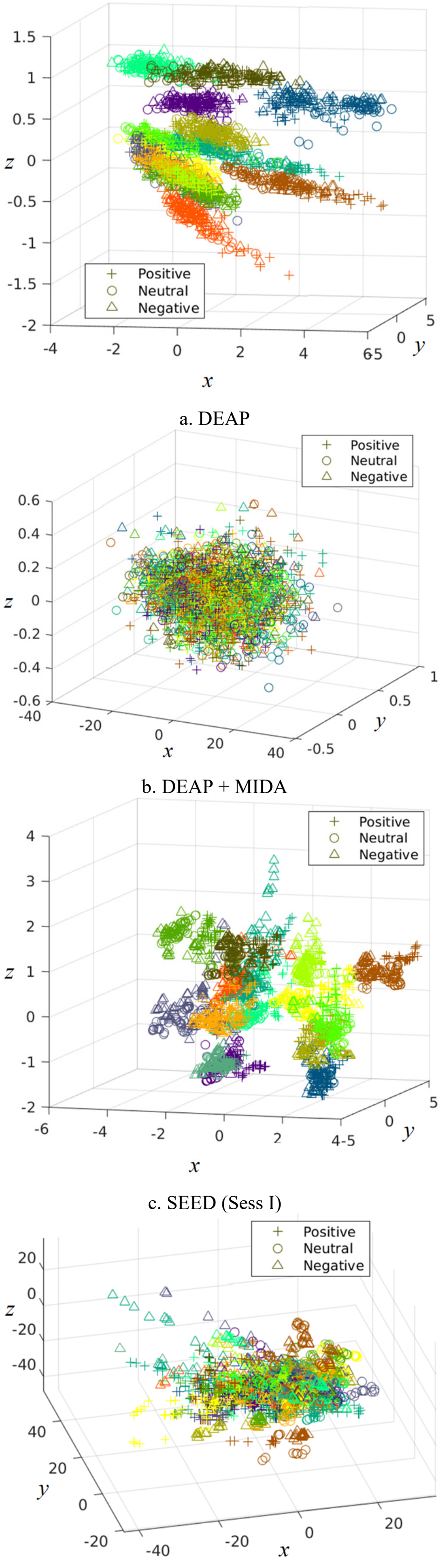
the upper bound of the chance level. The mean accuracy (std) of the baseline method on SEED is 54.73% (12.80%) on average over three sessions. The introduced domain adaptation methods can effectively enhance the mean classification accuracy up to 72.47% (t-test, $p < 0.05$). The best-performing methods are MIDA and TCA.

B. Cross-dataset domain adaptation

So far, current works [13-16] investigating domain adaptation methods in EEG-based emotion recognition have based their studies on one dataset: SEED. In the previous section, we present the study of domain adaptation methods on SEED as well as on another established dataset DEAP, and focus on the comparison between different domain adaptation techniques on the two datasets. In this section, we present a preliminary study of domain adaptation methods in a cross-dataset setting. Cross EEG dataset domain adaptation has not been addressed in the existing studies, and little is known about the performance of cross-dataset emotion classification. Conventionally, EEG studies have some constraints on the experiment settings. Notably, the training and test sessions adopt the same experiment paradigm, the same device or devices with the same technical specification. In our cross-dataset emotion classification experiment, the training data are contributed by one dataset and the test data by the other dataset. This experiment setting simulates the use case when the conventional setting could not be satisfied, and that the training data and test data are collected under different experimental paradigms using different EEG devices and affective stimuli. We stress that such investigation has been lacking thus far, and that it could make great practical sense as it relaxes the constraints on a conventional BCI.

We carry out six experiments to analyze the performance of cross-dataset emotion classification with and without using domain adaptation techniques: DEAP→SEED I, DEAP→SEED II, DEAP→SEED III, SEED I→DEAP, SEED II→DEAP, and SEED III→DEAP. The notation A→B denotes that dataset A contributes the training (source) data and dataset B contributes the test (target) data. SEED I, II and III denotes the data of session I, II and III from SEED, respectively. Since the domain adaptation methods require that the feature space of the source and target domain be the same, we use only the 32 channels in common between DEAP and SEED, and the DE features are 160-dimensional for both datasets. In the first three experiments, DEAP contributes the source training data containing $180 \times 14 = 2520$ samples. The classification performance is evaluated on a per-subject basis on SEED. The target test data contain 2775 samples. The mean classification performance is the average over 15 subjects in SEED. In the last three experiments, SEED contributes the source training data amounting to $2775 \times 15 = 41625$ samples. The classification performance is evaluated on a per-subject basis on DEAP. The test data contain 180 samples. The mean classification performance is the average over 14 subjects in DEAP. Due to limited computer memory, it is infeasible to include all source domain samples into the domain adaptation algorithms [13-14]. Therefore, we randomly sample 1/10 of the source data, amounting to 4162 samples, as actual source data for the domain adaptation algorithms. We then repeat the experiments 10 times and average the mean classification performance over 10 runs.

Table IV presents the results of the six cross-dataset experiments with different domain adaptation methods as well as with baseline method where no domain adaptation technique is used. As is shown in Table IV, the baseline accuracies range from 32.15% to 34.71%, and the values are below the upper bound of the 95 % confidence interval of the chance level. We therefore assert that the baseline performance is no significantly different from random guess at a 5 % significance level. It suggests that the technical differences between the two datasets may have introduced large discrepancies between the sample distributions of the source and target domains, besides the inter-



a. DEAP

b. DEAP + MIDA

c. SEED (Sess I)

d. SEED (Sess I) + MIDA

Fig. 1. Illustration of data sample distribution at feature level. Samples are reduced to 3 dimensions via Principal Component Analysis. The *x*, *y* and *z* axes correspond to the first, second and third principal components, respectively. +, ○ and Δ denote positive, neutral and negative samples, respectively. Colors represent different subjects. The perspective is adjusted to the best viewing angle for each plot. (a). DEAP, (b). DEAP + MIDA, (c). SEED (Sess I), (d). SEED (Sess I) + MIDA. (Best viewed in color.)

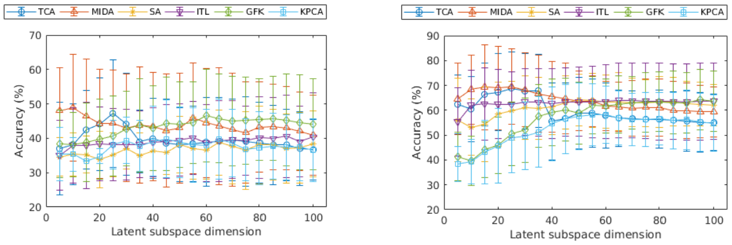
subject variance. Domain adaptation methods can effectively improve the accuracies over the baseline performance. TCA and MIDA are found to be the best-performing methods in the cross-dataset experiment settings: we observe 7.25 % – 13.40 % accuracy gains over the baseline performance.

V. Discussion

A. *Within-dataset domain adaptation*

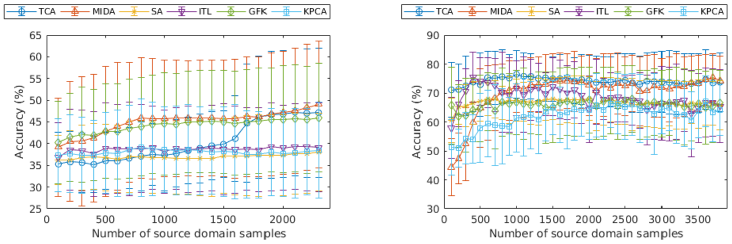
We present the study on the effectiveness of domain adaptation methods in a within-dataset leave-one-subject-out cross-validation setting. In this setting, each subject is hypothesized to constitute a domain by himself/herself, and domain discrepancy exists between different subjects. Several domain adaptation methods have been introduced to bridge the discrepancy between different subjects, so as to enhance the classification accuracy. In our study, domain adaptation methods work effectively on SEED, which coincides with the findings of [13-16]. MIDA and TCA are found to be the more effective methods, gaining an improvement of up to 20.66% over the baseline accuracy. On DEAP, domain adaptation could, to a less significant extent, improve the accuracy by up to 9.88%. We observe that domain adaptation methods work less effectively on DEAP than on SEED, which partially coincides with [16], which briefly mention that negative transfer had hindered the successful application of domain adaptation methods on DEAP. It remains an open question as to what determines the effectiveness of domain adaptation methods on a specific dataset. Here, we try to address this question with some empirical evidence. Fig. 1 presents the sample distribution of both datasets with and without using domain adaptation. As we can see in Fig. 1a and Fig. 1c, originally, the samples distribute differently between different subjects—each subject forms a cluster in the space by himself/herself. This suggests that large discrepancies between different subjects exist in the original feature space. We observe that samples are distributed more “orderly” on SEED than on DEAP. For example, on SEED, positive samples tend to locate on the right-hand side of each cluster. However, on DEAP we do not observe similar rules. In fact, samples belonging to different classes overlap substantially in each cluster, making it difficult to discriminate between different classes. Fig. 1b and Fig. 1d show the data sample distribution after applying MIDA. Clearly, the discrepancies between different subjects have been reduced, as the clusters are closer to each other. We observe that samples are better aligned on SEED then on DEAP. For example, on SEED, negative samples from different subjects tend to cluster in the upper space, while positive samples from different subjects tend to cluster in the lower space. However, on DEAP, samples of different classes are not well-aligned in the projected space. It might suggest that samples that are more “orderly” distributed in its original feature space tend to be better aligned in the domain invariant subspace, and it might explain why the baseline performance on SEED is superior to that on DEAP, and why domain adaptation methods give better performance on SEED than on DEAP.

DOI: 10.1109/TCDS.2018.2826840



a. DEAP b. SEED I

Fig. 2. Classification accuracy with varying latent subspace dimension on (a). DEAP and (b). SEED.



a. DEAP b. SEED I

Fig.3. Classification accuracy with varying number of source domain samples on (a). DEAP and (b). SEED.

1) *Latent dimension*

The dimension of the latent, domain-invariant subspace is one common hyperparameter among different domain adaptation methods. Fig. 2 presents the trend of mean classification accuracy with varying latent subspace dimension h. We observe that on both datasets, the performance of ITL is not so sensitive to varying h, but the other methods are. The best accuracies are obtained in a low-dimensional subspace, generally under 40. The optimal latent subspace dimension is considerably smaller than the original feature space dimension. It suggests that domain-invariant information may exist in a low dimensional manifold. Besides the benefits of domain invariance, a low dimensional latent space also reduces the burden of classifier training. Based on the finding, 10 – 30 is a suggested range for the latent subspace.

2) *Number of source samples*

Fig.3 presents the effect of varying number of source samples. On DEAP, the source dataset size varies from 100 to 2300. SA, ITL, and KPCA are less sensitive to varying number of source dataset size. For TCA, MIDA, and GFK, accuracies could be improved with growing number of source data. MIDA maintains a better accuracy than other methods at above 500 source domain samples. On SEED, the source dataset size varies from 100 to 3800. Similarly, accuracies are improved with more available source data. The accuracy flattens at above 1000 source domain samples. From that point onwards, MIDA and TCA perform similarly and are superior to the other methods. The finding suggests that if we have sufficient source data, MIDA and TCA tend to outperform other methods and hence the preferred techniques in terms of accuracy.

3) *Computation time*

The domain adaptation methods incur extra computational overhead. Table V shows the computation time for each domain adaptation method on both datasets. All experiments are simulated in MATLAB R2017a on a desktop PC equipped with one Intel Xeon E5-2630 CPU @ 2.20 GHz, 64 GB RAM, 512 GB SSD. On DEAP, $n_s = 2340$ and $n_t = 180$. On SEED, $n_s = 3885$ and $n_t = 2775$. The computation time is highest for ITL in both cases, due to its gradient-based iterative optimization. The two best-performing methods in terms of accuracy, TCA and MIDA, introduce considerable overheads. The major overheads of TCA, MIDA, and KPCA can be attributed to the eigendecomposition operation, which has a time complexity of $O(hn^2)$ [29]. This can become expensive when n grows to a large value. In existing studies [13-14, 17, 29, 31-32] simulated on averaged-specced PCs, n has been restricted to under 6000. Thus, they are more suitable for offline processing. The computation time of SA and GFK are almost negligible thanks to their closed-form solutions. SA and GFK might be used for online processing, but at the cost of lower accuracy performance.

B. *Cross-dataset domain adaptation*

We present a preliminary study of cross-dataset EEG-based emotion classification task. Conventionally, EEG-based applications have been constrained to using the same experiment protocol and device in the training and testing sessions. Clearly, it makes great practical sense if such constraints can be relaxed. In one scenario, for example, we could unite the high-quality datasets published by different research groups, and adapt those datasets to cater for our applicational need, instead of collecting and labeling new data from scratch. We set out to investigate the performance of cross-dataset emotion classification, where the two datasets are heterogeneous in various technical specifications, such as EEG devices, affective stimuli, and experiment protocol etc. We observe that the baseline accuracies without applying any domain adaptation method are performing at chance level. TCA and MIDA can effectively improve the classification performance over the baseline by 7.25 – 13.40%, suggesting that they could potentially reduce the technical discrepancies between datasets. However, though the accuracy improvements are significant (t-test, $p < 0.05$), the absolute accuracies remain below that of within-dataset training and testing. Considering the applicational values, more future studies on this topic are needed.

VI. Conclusion

In this paper, we present a comparative study on domain adaptation techniques on two affective EEG datasets, and a preliminary study on cross-dataset emotion recognition. We use two publicly available affective EEG datasets — DEAP and SEED. Though successful application of domain adaptation has

Table V. Computation time (s) of each domain adaptation method on both datasets.

	MIDA	TCA	SA	ITL	GFK	KPCA
DEAP	14.28	50.20	0.08	213.36	0.31	44.30
SEED	268.18	950.61	0.51	1348.70	1.33	992.14

DOI: 10.1109/TCDS.2018.2826840

been reported on SEED, little is known about the effectiveness of domain adaptation on other EEG datasets. We found that domain adaptation methods work more effectively on SEED than on DEAP. It remains an open question as to what determines the effectiveness of transfer learning techniques. The "orderliness" of the samples in the original feature space might have an impact on the effectiveness of adaptation.

The cross-dataset scheme simulates the use case where a conventional BCI paradigm cannot be satisfied. We demonstrate the effectiveness of MIDA and TCA in coping with domain discrepancy introduced by different subjects and the technical discrepancies with respect to the EEG devices, affective stimuli, experiment protocols etc. We stress that this is of great practical sense as it relaxes the constraint of a conventional BCI, but has been lacking sufficient investigation thus far. More future studies are needed on this topic.

References

[1] C. Mühl, B. Allison, A. Nijholt, and G. Chanel. "A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges," Brain-Computer Interfaces vol. 1, no. 2, pp. 66-84, May 2014

[2] R. Jenke, A. Peer, and M. Buss. "Feature extraction and selection for emotion recognition from EEG," IEEE Transactions on Affective Computing, vol. 5, no. 3, pp. 327–339, Sep. 2014.

[3] J. R. Wolpaw, D. J. McFarland, G. W. Neat, and C. A. Forneris. "An EEG-based brain-computer interface for cursor control," Electroencephalography and clinical neurophysiology, vol. 78, no. 3, pp. 252-259, Aug. 1990.

[4] J. R. Wolpaw, and D. J. McFarland. "Multichannel EEG-based brain-computer communication," Electroencephalography and clinical Neurophysiology, vol. 90, no. 6, pp. 444-449, Jun. 1994.

[5] T. Elbert, B. Rockstroh, W. Lutzenberger, and N. Birbaumer. "Biofeedback of slow cortical potentials. I," Electroencephalography and clinical neurophysiology, vol. 48, no. 3, pp. 293-301, 1980.

[6] N. Birbaumer, N. Ghanayim, T. Hinterberger, I. Iversen, B. Kotchoubey, A. Kübler, J. Perelmouter, E. Taub, and H. Flor. "A spelling device for the paralysed," Nature, vol. 398, no. 6725, pp. 297-298, Mar. 1999.

[7] M. Krauledat, M. Tangermann, B. Blankertz, and KR. Müller. "Towards zero training for brain-computer interfacing," PloS one, vol. 3, no. 8, e2967, Aug. 2008.

[8] S. Fazli, F. Popescu, M. Danóczy, B. Blankertz, KR. Müller, and C. Grozea. "Subject-independent mental state classification in single trials," Neural networks, vol. 22, no. 9, pp. 1305-1312, Nov. 2009.

[9] H. Kang, Y. Nam, and S. Choi. "Composite common spatial pattern for subject-to-subject transfer," IEEE Signal Processing Letters, vol. 16, no. 8, pp. 683-686, Aug. 2009.

[10] R. Lotte, and C. Guan. "Regularizing common spatial patterns to improve BCI designs: unified theory and new algorithms," IEEE Transactions on biomedical Engineering, vol. 58, no. 2, pp. 355-362, Feb. 2011.

[11] S. J. Pan, and Q. Yang. "A survey on transfer learning," IEEE Transactions on knowledge and data engineering, vol. 22, no. 10, pp. 1345-1359, Oct. 2010.

[12] V. Jayaram, M. Alamgir, Y. Altun, B. Scholkopf, and M. Grosse-Wentrup. "Transfer learning in brain-computer interfaces," IEEE Computational Intelligence Magazine, vol. 11, no. 1, pp. 20-31, Feb. 2016.

[13] W. L. Zheng, Y. Q. Zhang, J. Y. Zhu and B. L. Lu, "Transfer components between subjects for EEG-based emotion recognition," 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), Xi'an, 2015, pp. 917-922.

[14] W. L. Zheng, and B. L. Lu. "Personalizing EEG-based affective models with transfer learning," In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, 2016, pp. 2732-2738.

[15] X. Chai, Q. Wang, Y. Zhao, X. Liu, O. Bai, and Y. Li, "Unsupervised domain adaptation techniques based on auto-encoder for non-stationary EEG-based emotion recognition," Computers in biology and medicine, vol. 79, pp. 205-214, 2016

[16] X. Chai, Q. Wang, Y. Zhao, Y. Li, D. Liu, X. Liu, and O. Bai. "A Fast, Efficient Domain Adaptation Technique for Cross-Domain Electroencephalography (EEG)-Based Emotion Recognition," Sensors vol. 17, no. 5, pp. 1014, 2017

[17] K. Yan; L. Kou; D. Zhang, "Learning Domain-Invariant Subspace Using Domain Features and Independence Maximization," in IEEE Transactions on Cybernetics, in print.

[18] L. C. Shi, Y. Y. Jiao and B. L. Lu, "Differential entropy feature for EEG-based vigilance estimation," 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Osaka, 2013, pp. 6627-6630.

[19] S. Koelstra, C. Muhl, M. Soleymani, JS. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, "DEAP: A Database for Emotion Analysis Using Physiological Signals," IEEE Transactions on Affective Computing, vol. 3, no. 1, pp. 18-31, Mar. 2012.

[20] W. L. Zheng and B. L. Lu, "Investigating Critical Frequency Bands and Channels for EEG-Based Emotion Recognition with Deep Neural Networks," IEEE Transactions on Autonomous Mental Development, vol. 7, no. 3, pp. 162-175, Sep. 2015.

[21] T. Higuchi. "Approach to an irregular time series on the basis of the fractal theory," Physica D: Nonlinear Phenomena, vol. 31, no. 2, pp. 277-283, Jun. 1988.

[22] R. W. Picard, E. Vyzas, and J. Healey. "Toward machine emotional intelligence: Analysis of affective physiological state," IEEE transactions on pattern analysis and machine intelligence, vol. 23, no. 10, pp. 1175-1191, Oct. 2001.

[23] Y. Liu, and O. Sourina. "Real-time subject-dependent EEG-based emotion recognition algorithm," Transactions on Computational Science XXIII, pp. 199-223. 2014.

[24] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. "Measuring statistical dependence with Hilbert-Schmidt norms," In International conference on algorithmic learning theory, Springer Berlin Heidelberg, pp. 63-77, Oct. 2005.

[25] C. W. Seah, Y. S. Ong, and I. W. Tsang. "Combating negative transfer from predictive distribution differences," IEEE transactions on cybernetics vol. 43, no. 4, pp. 1153-1165, Aug. 2013.

[26] B. Schölkopf, A. Smola, KR. Müller. "Nonlinear component analysis as a kernel eigenvalue problem," Neural computation, vol. 10, no. 5, pp. 1299-1319, Jul. 1998.

[27] S. Mika, G. Ratsch, J. Weston, B. Scholkopf and K. R. Mullers, "Fisher discriminant analysis with kernels," Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop, Madison, WI, 1999, pp. 41-48.

[28] L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt. "Feature selection via dependence maximization," Journal of Machine Learning Research, vol. 13, pp. 1393-1434, May. 2012.

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

[30] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 2960-2967.

[31] Y. Shi and F. Sha, "Information-theoretical learning of discriminative clusters for unsupervised domain adaptation," in Proceedings of the Intl. Conf. on Machine Learning (ICML), 2012.

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

[33] B. Schölkopf, A. Smola, and K. R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem. Neural computation,", vol. 10, no. 5, pp. 1299-1319, Jul. 1998

[34] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola, "A kernel method for the two-sample problem," in Proc. Conf. Neural Inf. Process. Syst. 19. Cambridge, MA, 2007, pp. 513–520.

[35] A. J. Smola, A. Gretton, L. Song, and B. Schölkopf, "A Hilbert space embedding for distributions," in Proc. 18th Int. Conf. Algorithmic Learn. Theory, Sendai, Japan, Oct. 2007, pp. 13–31.

[36] G. Müller-Putz, R. Scherer, C. Brunner, R. Leeb, and G. Pfurtscheller, "Better than random: a closer look on BCI results," International Journal of Bioelectromagnetism, vol. 10, pp. 52-55, 2008.

[37] E. Combrisson and K. Jerbi, "Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy," Journal of Neuroscience Methods, vol. 250, pp. 126-136, 2015.