**Proceedings Paper:**

Giles, J., Ang, K.K., Mihaylova, L. et al. (1 more author) (2020) Weighted transfer learning of dynamic time warped data for motor imagery based brain computer interfaces. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 42nd Annual International Conferences of the IEEE Engineering in Medicine and Biology Society (EMBC 2020), 20-24 Jul 2020, Online conference. IEEE . ISBN 9781728119915

https://doi.org/10.1109/EMBC44109.2020.9176635

# Weighted Transfer Learning of Dynamic Time Warped Data for Motor Imagery based Brain Computer Interfaces

Joshua Giles[1], Kai Keng Ang[2], Lyudmila Mihaylova[1], Mahnaz Arvaneh[1]

***Abstract*— A large amount of calibration data is typically needed to train an electroencephalogram (EEG)-based brain-computer interfaces (BCI) due to the non-stationary nature of EEG data. This paper proposes a novel weighted transfer learning algorithm using a dynamic time warping (DTW) based alignment method to alleviate this need by using data from other subjects. DTW-based alignment is first applied to reduce the temporal variations between a specific subject data and the transfer learning data from other subjects. Next, similarity is measured using Kullback Leibler divergence (KL) between the DTW aligned data and the specific subject data. The other subjects' data are then weighted based on their KL similarity to each trials of the specific subject data. This weighted data from other subjects are then used to train the BCI model of the specific subject. An experiment was performed on publicly available BCI Competition IV dataset 2a. The proposed algorithm yielded an average improvement of 9% compared to a subject-specific BCI model trained with 4 trials, and the results yielded an average improvement of 10% compared to naive transfer learning. Overall, the proposed DTW-aligned KL weighted transfer learning algorithm show promise to alleviate the need of large amount of calibration data by using only a short calibration data.**

## I. INTRODUCTION

Brain computer interfaces (BCI) are devices allowing for communication between a person and machine through signals collected from the brain directly [1], [2]. Recently more is being invested into BCI by large companies as the potential for both gaming and medical applications is recognised [1]. Electroencephalogram (EEG)-based BCI are of particular interest due to its high temporal resolution while being non-invasive [1].

Despite EEG obtaining high temporal resolution, a long calibration session is required to retrain the feature extractor and classifier of the BCI before each use [3]. This is due to the non-stationary nature of the EEG signals being collected. One approach to reduce the need for a long calibration period is through transfer learning [4]. Transfer learning improves training of a new model by transferring knowledge collected from other related tasks [4]. In BCI transfer learning is commonly performed between either different sessions from the same user or different users completing the same task. Doing this naively, by pooling all the data without any weighting, removes the need for a calibration session however results in a deteriorated classification accuracy.

To improve the accuracy, a small number of subject-specific labelled trials can be collected and used to adjust the transfer learning algorithm. In session to session transfer learning the labelled trials allow the application of data space adaptation (DSA) [5] [6]. This utilises a linear transform to reduce the Kullback Leibler divergence between the testing and training data sets. The use of these subject-specific labelled trials has also been explored in subject to subject transfer learning in a number of studies [7]–[9]. For example Lotte *et al* used the labelled trials to regularise the transferred data before training the BCI [7].

Another approach that has been explored to improve the classification accuracy is to identify the users who will benefit from transfer learning. The transfer learning is then only applied to the subjects that require it.The proposed Jensen Shannon Ratio Framework belongs to this approach and showed success to some extent [10]. Overall there are still a number of limitations with the available algorithms belonging to this approach. They either require a large amount of past data from the same subject or struggle to apply the data effectively to build a model for the new user.

The novel algorithm proposed in this paper, referred to as DTW KL weighted TL, uses a limited number of subject-specific trials to adjust and weight the transferred data. Initially our proposed algorithm uses dynamic time warping (DTW) to align the transfer learning data from every other subject [11]. This alignment reduces the temporal variations between each subject's transfer learning data and the subject-specific data. Afterwards the Kullback Leibler (KL) divergence is used to measure the similarity between each subject's aligned transfer data and the subject-specific data. This KL value is then used to weight the aligned data sets before using them to train the feature extractor and the classifier of the BCI.

The publicly available BCI Competition IV data set 2a [12] will be used to evaluate the proposed DTW KL weighted algorithm. The results will be compared with the classification results of four other algorithms. First a naive transfer learning algorithm, where the BCI model is trained using a pool of data from other subjects without any weighting and alignment (i.e. Naive TL). The second algorithm trains the BCI model using a pool of data from other subjects after aligning to the few subject-specific data without any weighting (i.e. DTW Naive TL). The third algorithm uses the weighted pool data of other subject without alignment (i.e. KL weighted TL). Finally a state of the art subject-

* This work was supported by University of Sheffield, UK and the Agency for Science, Technology and Research (A*STAR), Singapore.

[1] J. Giles, L. Mihaylova, and M. Arvaneh are with Automatic control and System Engineering Department, University of Sheffield , Sheffield, UK (emails: jgiles1, l.s.mihaylova, m.arvaneh@sheffield.ac.uk).

[2] K. K. Ang is with Institute for Infocomm Research,Agency for Science, Technology and Research (A*STAR), Singapore 138632. (email: kkang@i2r.a-star.edu.sg).

specific algorithm where the BCI model is trained with the few available subject-specific data.

## II. Methodology

### A. Dynamic Time Warping-based alignment

Dynamic Time Warping (DTW) is commonly used to reduce temporal variations between two time series. The proposed algorithm uses a DTW-based alignment method to transform the data from other subjects in the time domain such that each trial of the other subjects is aligned to the averaged trials from the same class in the specific subject. These transformed trials are then used as to find the average DTW aligned covariance matrix to represent the transfer learning data.

Representing each other subject's EEG transfer learning data as $\hat{\mathrm{D}} = (\hat{\mathbf{X}}_i, \hat{y}_i)_{i=1}^{N}$, the $i^{th}$ trial $\hat{\mathbf{X}}_i \in \hat{\mathbf{X}} \subset R^{ch\times t}$ is the recorded data with $ch$ being the number of channels and $t$ representing the number of time samples. $\hat{y}_i \in \hat{\mathbf{y}} \subset \mathbb{R}$ represents the corresponding class label. Similarly, the subject-specific data is collected in a short calibration session from the new user, containing a small number of trials. For this data $D = (\mathbf{X}_i, y_i)_{i=1}^{N_l}$ where $\mathbf{X}_i \in \mathbf{X} \subset R^{ch\times t}$ is the $i^{th}$ recorded trial and $y_i \in \mathbf{y} \subset \mathbb{R}$ represents its corresponding class label. The average subject-specific trial is calculated through (1) with $l_c$ labelled trials being available, of class $c$.

$$\overline{\mathbf{X}}_c = \frac{1}{l_c}\sum_{i=1}^{l_c}\mathbf{X}_c^i \tag{1}$$

DTW-based alignment is used on each of the transfer learning trials of the other subjects, aligning them to the average of the subject-specific trials of the same class. In order to align the trials through DTW a $t-by-t$ distance matrix, $\mathbf{D}$, is created measuring the Euclidean distance between each time point, Where $\mathbf{D}(a,b)$ of the matrix present the distance between time instances $a$ and $b$ from $\hat{\mathbf{X}}_c$ and $\overline{\mathbf{X}_\mathbf{c}}$ respectively averaged across all the available channels.

$\mathbf{W}$ is a warping path whose elements define a mapping between $\hat{\mathbf{X}}_\mathbf{i}^\mathbf{c}$ and $\overline{\mathbf{X}}^\mathbf{c}$. The $k_{th}$ element of $\mathbf{W}$ is defined as: $w(k) = \mathbf{D}(a_k, b_k)$ therefore:

$$\mathbf{W} = [w(1),..,w(k),..,w(K)], \quad t \le K < 2t-1 \tag{2}$$

This warping path is subject to a number of constraints including the boundary conditions: $w(1) = \mathbf{D}(1,1)$ and $w(K) = \mathbf{D}(t,t)$, as well as the continuity constraint and the monotonicity constraint: $0 \le a_K - a_{K-1} \le 1$ and $0 \le b_K - b_{K-1} \le 1$. The optimum warping path, $\mathbf{W}^*$, is selected to minimize the alignment between $\hat{\mathbf{X}}_\mathbf{c}$ and $\overline{\mathbf{X}}_\mathbf{c}$,

$$\mathbf{W}^* = min(\frac{1}{k}\sum_{k=1}^{K} w(k)). \tag{3}$$

This produces a newly aligned trial shown in (4) where $[a_1^*, ..., a_K^*]$ are the time indices of $\hat{\mathbf{X}}_\mathbf{c}^\mathbf{s}$, and $s$ refers to the transfer learning subject number.

$$\hat{\mathbf{X}}^s{}_{c_{aligned}} = \begin{bmatrix} \hat{\mathbf{X}}_c^s(a_1^*,1) & ... & \hat{\mathbf{X}}_c^s(a_1^*,ch) \\ ... & ... & ... \\ \hat{\mathbf{X}}_c^s(a_K^*,1) & ... & \hat{\mathbf{X}}_c^s(a_K^*,ch) \end{bmatrix} \tag{4}$$

These aligned trials are then used to calculate the covariance of the data using (5), with the $l_c$ labelled trials available. The aligned covariance can then be used to train the BCI.

$$\hat{\mathbf{\Sigma}}^s_{c_{aligned}} = \frac{1}{l_c}\sum_{i=1}^{l_c}\frac{\hat{\mathbf{X}}^{s^i}_{c_{aligned}}\hat{\mathbf{X}}^{s^{iT}}_{c_{aligned}}}{tr(\hat{\mathbf{X}}^{s^i}_{c_{aligned}}\hat{\mathbf{X}}^{s^{iT}}_{c_{aligned}})} \tag{5}$$

### B. Weighted Transfer Learning

The second step of the proposed algorithm takes the aligned transfer data of every other subject and then weights the data based on its similarity to the subject specific data. The KL divergence is then used to calculate these weights for every other subject's aligned covariance before they are combined and used to train the common spatial patterns (CSP) for the BCI.

Applying a weighting to every other subject's transfer learning data reduces the effects any of the irrelevant data has on training the BCI, while the effective subjects are focused on. The weighting is calculated based on the similarity between the limited subject-specific data collected and each of the other subjects transfer learning data. These sets of EEG data can be modelled as Gaussian distributions with a mean of zero, after band pass filtering, and the covariance calculated as shown in (6), using the mean trial data calculated in (1). As such the Kullback Leibler divergence can be used to measure their similarity.

$$\mathbf{\Sigma} = \frac{\mathbf{X}\mathbf{X}^T}{tr(\mathbf{X}\mathbf{X}^T)} \tag{6}$$

Kullback Leibler divergence is a commonly used measure between two Gaussian distributions. Given two Gaussian distributions $N_0(\mu, \Sigma)$ and $N_1(\hat{\mu}, \hat{\Sigma})$ with $\hat{\mu}$ and $\mu$ representing the means of the distribution while $\hat{\mathbf{\Sigma}}$ and $\mathbf{\Sigma}$ co-variances. The KL divergence between the two Gaussian distributions can be calculated using (7),

$$\begin{aligned} KL[N_0 \parallel N_1] = \frac{1}{2}[(\hat{\mu}-\mu)^T\hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}-\mu) \\ +tr(\hat{\mathbf{\Sigma}}^{-1}\mathbf{\Sigma}) - ln(\frac{det(\mathbf{\Sigma})}{det(\hat{\mathbf{\Sigma}})}) - k]. \end{aligned} \tag{7}$$

As the band pass filtered EEG signals have a mean of 0, so the only focus is on the covariance of the signal. In this proposed algorithm the KL between the mean subject specific data, $\overline{\mathbf{X}}_\mathbf{c}$, and every other subjects aligned transfer learning data, $\hat{\mathbf{X}}_c^{aligned}$, for each of the classes is calculated. A mean KL of the classes is then calculated for each of the transfer learning subject through equation (8), where $s$ refers to the subject number.

$$KL_s = \sum_{c=1}^{C} KL[\overline{\mathbf{X}}_c] \parallel \hat{\mathbf{X}}^s{}_{c_{aligned}}] \tag{8}$$

The weighting for each transfer learning subject is computed through equation (9) and (10). This maximises the weighting for subjects with similar data to the calibration data while subjects with no similarity are given a small weighting.

$$OE_s = \frac{\sum_{s=1}^{S} KL_s}{KL_s} \tag{9}$$

$$We_s = \frac{OE_s}{\sum_{s=1}^{S} OE_s} \tag{10}$$

$$\mathbf{\Sigma}_{tl} = \frac{\sum_{s=1}^{S} \hat{\mathbf{\Sigma}}^{s}{}_{aligned} We_s}{S} \tag{11}$$

These weightings are then combined with each subjects covariance to calculate a transfer learning covariance using equation (11). This transfer learning covariance is used for training a CSP filter. A linear discriminant analysis classifier is then trained using the CSP.

## III. Experiment

### A. Data set

The publicly available "Graz data set A" from BCI Competition 2008 [12] was used in this paper. This data sets consists of EEG data of nine subjects performing four classes of motor imagery, however only left and right hand motor imagery were used for evaluation of the proposed algorithm as it focuses on binary class BCI. For each subject, the EEG data from only the first session, consisting 144 trials of right and left hand motor imagery, was used in this study.

### B. Data processing

The performance of the proposed DTW KL weighted algorithm was compared with the results of four other algorithms, namely subject-specific, Naive TL, DTW Naive TL and KL weighted TL. The first 40 trials of the subject-specific data was set aside for calculating the DTW and the weighting for every other subjects data. The data was band pass filtered with a zero phase elliptic filter from 8 Hz to 35 Hz before being put into the algorithms.

Thereafter, the proposed DTW KL weighted algorithm used the KL weighted aligned data from the other subjects to train the CSP filters, while the pooled previous data without alignment and weighting were used for training of the CSP filters in the Naive TL algorithm. Similarly, DTW Naive TL pooled aligned data from the other subjects, and KL weighted TL pooled the weighted data from the other subjects to train the CSP filters. The subject-specific algorithm used only the 40 subject-specific trials to train the CSP filters. Finally the CSP features of the subject-specific trials were used to train an LDA classifier.

## IV. Results and Discussion

As can be seen in Fig. 1, the implementation of transfer learning, including Naive, DTW Naive and KL weighted, can produce large increases in classification accuracy for some subjects. Figure 1 shows how implementing transfer learning increases the classification accuracy for subject 2 when there are four trials available. For this subject, using Naive TL improves the accuracy by 9.6% while KL weighted TL increases the accuracy by 15%. This improvement is also apparent for subjects 1,3 and 8, however on average a better accuracy is achieved through relying on the subject-specific trials when compared with the average results of Naive TL, DTW Naive TL and KL weighted TL.

Using weights on the transfer learning data limits the effects of irrelevant data when training the BCI which leads to a more representative model for the test subject.Subject 1 in particular benefits from weighting the transfer learning. After a 13.5% drop from naive transfer learning the KL weighted transfer learning achieves an improvement of 12.4%. This change of 26% highlights the impact that utilising appropriate weightings for the data can have. Some other subjects experience a fall in accuracy once their data is weighted, such as subject 6. This fall in accuracy highlights that relying on KL weightings alone cannot always provide the appropriate weightings, particularly in a small data set. To improve this the similarity between the transfer learning data sets and the test data has to be increased using the proposed DTW KL weighted TL.

Overall the mixed changes in accuracy across the different subjects lead to an average fall of accuracy of 3% when naive transfer learning was implemented compared to the subject specific BCI. While the KL weighted transfer learning performed slightly better with an average decrease of 0.6%. Due to this, neither of the transfer learning algorithms show a statistically significant difference from relying on the subject-specific data. This average drop in accuracy could be due to the limited transfer learning pool. A larger pool of subjects could allow the algorithm to find data closer to the test subjects.

Despite this fall in the average accuracy for a number of subjects, there is a significant improvement in accuracy. To take advantage of these improvements and make the improvement statistically relevant the KL weighted transfer learning needs to be made more generalizable. Applying DTW to reduce the time variance between the transfer learning data and subject-specific data should help accomplish this.

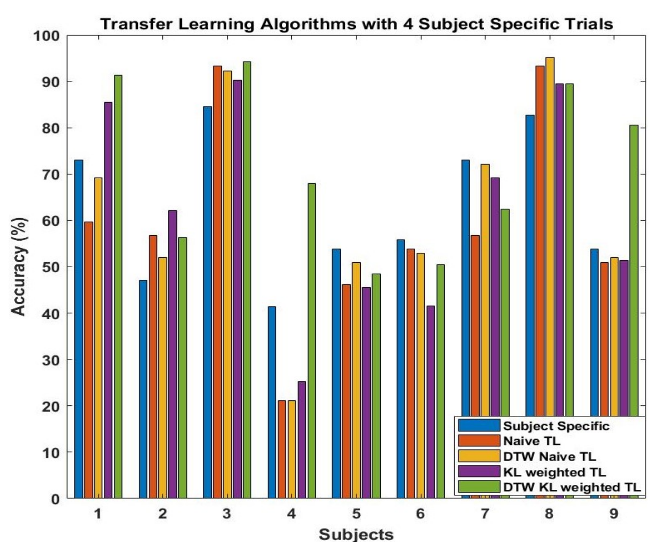


Fig. 1. The accuracy achieved by each subject through transfer learning when only 4 subject-specific trials are collected for calibrating the BCI

The DTW aligned transfer learning data has an increase in similarity to the subject specific data due to the reduced time variance. Stretching the transfer learning data in the time domain allows for a more accurate calculation of the similarity between the subject-specific data and transfer learning dataFigure 1 shows the effects of DTW on both naive and KL transfer learning. There is no significant effect of DTW on naive transfer learning, only a minor improvement in the average accuracy of 2%. Despite the time variance being reduced through DTW without any weighting, detrimental data still limits the effectiveness of the transfer learning. Although this detrimental data is now time aligned if the activation pattern is completely different the effectiveness of reducing the time variance is limited.

There is a significant improvement in accuracy when DTW is combined with the KL weighted transfer learning (i.e. our proposed DTW KL weighted TL) with an average improvement of 7.2%. Using DTW to align the transfer learning data before the similarity to the test subject is measured allows a more accurate comparison between the data sets. A more accurate comparison improves the weightings calculated and the model being trained using the weighted transfer learning data. Subject 9 in particular achieves a 46% increase in accuracy using the proposed DTW KL weighted TL compared to the accuracy obtained by the KL weighted TL.

As more subject-specific trials are collected, the proposed DTW-aligned KL weighted transfer learning algorithm continues to outperform the average subject-specific trained BCI (see Fig. 2). This average improvement in accuracy, despite the use of the same data for training (i.e. only data from other subjects), is due to the more accurate estimations of the similarity through KL divergence. The difference between the subject-specific accuracy and the proposed DTW KL weighted TL slowly decreases as more trials are available making the subject specific BCI gets more robust. Thus, as shown in Fig. 2, when the subject-specific trials increases from 4 to 20, the average improvement obtained using the proposed DTW KL weighted TL gets limited from 7.2% to 3%.

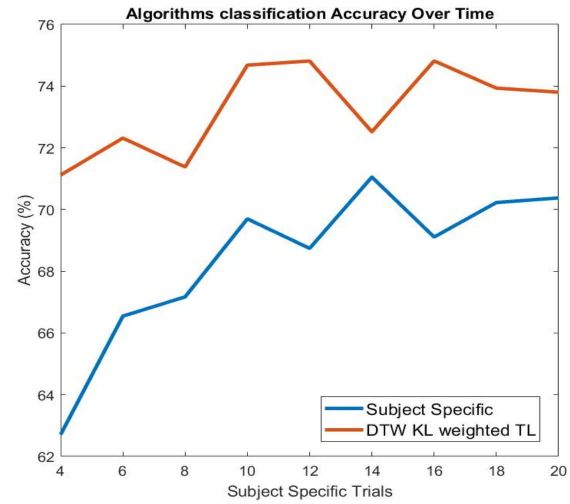


Fig. 2. Comparison between the average accuracy of the proposed algorithm (DTW KL weighted TL) and the average subject-specific accuracy, across different number of subject-specific trials available for calibration

## V. Conclusion

Overall the proposed DTW-aligned KL weighted transfer learning algorithm was able to improve the accuracy with only a short calibration period, significantly outperforming both naive transfer learning and subject specific BCI. The DTW based alignment successfully reduces the temporal non-stationary between the trials of the new subject and previously recorded trials, and KL weighting reduces the impact of data with less similarity with the subject-specific trials. Thus, the proposed algorithm outperformed the standard BCI by 9% when there are only four subject specific trials available.

Although the proposed algorithm leads to an average improvement in accuracy there are still some subjects that perform better when relying on subject- specific data to train the BCI. This could be due to the limited transfer subject pool of only 9 subjects. A larger pool of subjects would allow the algorithm to find more subjects close to the test subject and utilise them. Evaluating this algorithm on a data set with more subjects may lead to improvements in accuracy as closer transfer learning subjects are made available.

## References


[1] T. W. Berger, J. K. Chapin, G. A. Gerhardt, D. J. McFarland, J. C. Principe, W. V. Soussou, D. M. Taylor, and P. A. Tresco, *Brain-Computer Interfaces*. Dordrecht: Springer Netherlands, 2008.

[2] K. K. Ang, S. Member, C. Guan, and S. Member, "EEG-Based Strategies to Detect Motor Imagery for Control and Rehabilitation," *IEEE Trans Neural Syst Rehabil Eng*, vol. 25, no. 4, pp. 392–401, 2017.

[3] C. Vidaurre, C. Sannelli, and M. Klaus-robert, "Co-adaptive calibration to improve BCI," *J. Neural Eng.*, vol. 8, 2011.

[4] A. M. Azab, J. Toth, L. S. Mihaylova, and M. Dr Arvaneh, "A review on transfer learning approaches in brain–computer interface," in *Signal Processing and Machine Learning for Brain-Computer Interfaces*, ch. 5, pp. 96 – 116, 2018.

[5] M. Arvaneh, I. Robertson, and T. E. Ward, "Subject-to-subject adaptation to reduce calibration time in motor imagery-based brain-computer interface," in *EMBC 2014*, pp. 6501–6504, 2014.

[6] J. Giles, K. Ang, L. Mihaylova, and M. Arvaneh, "Data Space Adaptation for Multiclass Motor Imagery-based BCI," in *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, IEEE, apr 2018.

[7] F. Lotte and C. Guan, "Regularizing common spatial patterns to improve BCI designs: Unified theory and new algorithms," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 2, pp. 355–362, 2011.

[8] H. Kang, Y. Nam, and S. Choi, "Composite Common Spatial Patterns for Subject-to-Subject Transfer," *IEEE Signal Process. Lett.*, vol. 16, no. 8, pp. 683–686, 2009.

[9] H. Kang and S. Choi, "Bayesian common spatial patterns for multi-subject EEG classification," *Neural Netw.*, vol. 57, pp. 39–50, 2014.

[10] J. Giles, K. K. Ang, L. S. Mihaylova, and M. Arvaneh, "A Subject-to-subject Transfer Learning Framework Based on Jensen-Shannon Divergence for Improving Brain-computer Interface," in *IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 3087 – 3091, 2019.

[11] A. M. Azab, L. Mihaylova, H. Ahmadi, and M. Arvaneh, "Robust Common Spatial Patterns Estimation Using Dynamic Time Warping to Improve BCI Systems," in *IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019.

[12] C. Brunner, R. Leeb, G. R. Muller-Putz, A. Schlogl, and G. Pfurtscheller, "BCI Competition 2008 - Graz data set A," 2008.