Transfer Learning to Account for Idiosyncrasy in Face and Body Expressions

Bernardino Romero-Paredes1, Hane Aung1, Massimiliano Pontil2, Amanda C. de C. Williams3, Paul Watson4, Nadia Bianchi-Berthouze5

1 Dept. of Computer Science, UCL Interactive Centre, University College London, UK
2 Dept. of Computer Science, University College London, UK
3 Clinical, Educational, and Health Psychology, University College London, UK
4 Department of Health Sciences, Leicester General Hospital, University of Leicester, UK
5 UCL Interactive Centre, Division of Psychology & Language Sciences, University College London, UK

Abstract—In this paper we investigate the use of the Transfer Learning (TL) framework to extract the commonalities across a set of subjects and also learns the way each individual instantiates these commonalities to model idiosyncrasy. To implement this we apply three variants of the Multi Task Learning, namely: Regularized Multi Task Learning (RMTL), Multi Task Feature Learning (MTFL) and Composite Multi Task Feature Learning (CMTFL). Two datasets are used; the first is a set of point based facial expressions with annotated discrete levels of pain. The second consists of full body motion capture data taken from subjects diagnosed with chronic lower back pain. A synchronized electromyographic signal from the lumbar para-spinal muscles is taken as a pain-related behavioural indicator. We compare our approaches with Ridge Regression which is a comparable model without the Transfer Learning property; as well as with a subtractive method for removing idiosyncrasy. The TL based methods show statistically significant improvements in correlation coefficients between predicted model outcomes and the target values compared to baseline models. In particular RMTL consistently outperforms all other methods; a paired t-test between RMTL and the best performing baseline method returned a maximum p-value of $2.3 \times 10^{-4}$.

I. INTRODUCTION

Face and body movements are highly informative channels to communicate and infer emotional state and intensity. However, these movement patterns often inherently contain idiosyncrasies. These are further compound with artifacts stemming from variable environmental influences. These factors pose particular challenges in developing recognition models which seek to identify behaviours and expressions attributed to psychological states and affect in people. This confound is particularly salient if available datasets do not contain a high number or a wide enough variety of subjects. In such cases the biases caused by idiosyncratic behaviours could be significant if it is ignored or not treated carefully. This is particularly important for researchers focusing on naturalistic expressions. In many affect recognition studies this issue is not taken as a central consideration. (For a comprehensive survey which includes affect recognition from the face see [26] and specifically for body motion, see [13].

Additionally Gao et al. reviews methods to automatically recognise affect from touch behaviour [11]).

Bernhardt & Robinson [4] showed that removal of idiosyncrasy improved the performance of Support Vector Machine based models in classifying affect from hand and arm motions. In that study, idiosyncrasy was considered to be an additive constant and individuals’ mean feature values were subtracted as a way to remove this bias. [24] apply affine based registration obtained using calibration samples to remove inter-personal shape differences for the recognition of temporal phases in facial expression. A different approach was proposed by [16] which decouples the contribution caused by identity and expression using bilinear models [22]. This strategy is interesting for analyzing and synthesizing different combinations of factors (identity-affect) but at present in supervised learning it is constrained to classification and does not generalize to regression.

The aim in this study is not to decouple and subtract identity salient information and apply models to the remainder. Rather the view is taken that there exists a commonality within a group of subjects expressing the same affective state, and that each subject will provide an instance of this commonality in a different way. It is the differences in these instantiations that represents idiosyncrasy and we seek to use this information rather than subtract it. For example, during periods of intense pain, the shrinking of the eye can be considered to be a commonality but each individual will give an instance of this in their own idiosyncratic way. To this end we exploit methodologies that can determine the commonalities but are also adaptive to new instances. One such paradigm which has this capacity is Transfer Learning (TL) [17] (also known as Learning to Learn [14] or Inductive Transfer [8, 19, 23]). In general TL extracts knowledge from a set of transfer tasks to be applied in the learning process of one or more target tasks. It is inspired by the fact that human beings are able to use knowledge from previous learned tasks in order to face a new one [3]. There are works in psychology [5] which provide support to the hypothesis that the knowledge acquired by human beings in learning physical movements makes easier the learning process of new movements.
In this study, we investigate the implementation of three variants of MTL which we will refer to as 1) Regularized Multi Task Learning (RMTL), 2) Multi Task Feature Learning (MTFL) and 3) Composite Multi Task Feature Learning (CMTFL). Section II will further describe these three proposed methodologies outlining the differences and the assumptions. Section III details the acquisition methods and content of the two datasets used in this study. The results section IV shows comparative correlation coefficients between the predicted outputs of each of the proposed models as well as three baseline models and the target outcomes. Lastly, we conclude and discuss the findings in section V.

II. METHODS FOR TRANSFER LEARNING ON FACE AND BODY MOTION

In this study three variants of MTL are implemented which are based on optimization problems that use regularizers on the task parameters to encode the relationships among the tasks. For all three cases, learning is applied over two stages: a transfer stage and a calibration stage. In the transfer stage, data from a group of subjects (which we will refer to as transfer subjects) performing actions which can convey information about a ground truth is gathered. A supervised learning model is trained to extract all of the information about a ground truth is gathered. A supervised learning model is trained to extract all of the information.

A. Notation

We assume the following setting: the model we want to train has to be able to learn a set of $T$ transfer linear tasks $f_t(x) = \langle w_t, x \rangle$, $x, w_t \in \mathbb{R}^d$, $\forall t \in \{1 \ldots T\}$, where $d$ is the dimensionality of the data. In order to learn these tasks, a set of labeled instances is provided: $\{X_t, Y_t\}$, $X_t \in \mathbb{R}^{d \times m_t}$, $Y_t \in \mathbb{R}^{m_t}$, where $X_t$ is the matrix composed of all the $m_t$ instances provided for task $t$ as columns and $Y_t$ contains the labels. With a slight abuse of notation, we denote $X = \{X_1, X_2, \ldots, X_T\}$ and $Y = \{Y_1, Y_2, \ldots, Y_T\}$. Let $W$ be the matrix composed of the $T$ weight vectors $w_t$ as columns, that is, $W = [w_1, \ldots, w_T]$. Then in the transfer stage the following optimization problem is solved:

$$\min_{W,C} \sum_{t=1}^{T} \|X_t^\top w_t - Y_t\|_2^2 + \gamma \Omega(W, C), \quad (II.1)$$

where $W$ contains the estimators for the $T$ transfer tasks and $C$ contains information about the commonalities among the tasks. $\Omega$ is called the regularizer and provides an intuitive mechanism to incorporate similarities among tasks weight vectors; and $\gamma > 0$ is a hyperparameter which needs to be tuned beforehand and ponders the importance of the regularizer with respect to the empirical loss. As noted above, we will consider three different functions $\Omega$, which implement three different sets of assumptions about the relationships between the tasks. We focus on regression problems but the generalization to classification problems is straightforward by changing the loss function.

We denote as $X_C$ and $Y_C$ the instances and labels provided by the target subject for the calibration stage. To do so, we need to take $C$ and optimize the following problem:

$$\min_{w_C} \sum_{t=1}^{T} \|X_C^\top w_C - Y_C\|_2^2 + \gamma \Omega(w_C, C). \quad (II.2)$$

Note that in this case $C$ is fixed and the optimization is performed only on the weight vector of the target task $w_C$. Once this weight vector $w_C$ is learned, it can be used to perform inference on new instances. In the following we will study three MTL approaches which can be described within this framework and use different regularizers $\Omega$ and different ways to encode the common information $C$ learned in the transfer stage.

B. Regularized MTL

One way of defining relations among a set of tasks is by looking at similarities between the parameters. Given that the tasks are defined by linear functions, one may assume that their weight vectors are close to each other; therefore, one could penalize the distances among these weight vectors. One way of formulating this idea is by assuming that all tasks weight vectors are close to a reference vector $w_0$ which also needs to be learned. This idea was proposed in [9], where the authors assume that $w_t = w_0 + v_t$, $\forall t \in \{1 \ldots T\}$,
where $v_t$ is what makes task $t$ different from the others and therefore it is assumed to be small. This assumption is taken into account by adding regularization terms into the problem which penalizes the square distance between any $w_t$ and $w_0$ (in other words, penalizing the module of $v_t$, $\forall t \in \{1 \ldots T\}$).

The resultant approach is based on the following regularizer:

$$
\Omega(W, w_0) = \frac{1}{T} \sum_{t=1}^{T} \|w_t - w_0\|^2 + \lambda \|w_0\|^2, \quad (II.3)
$$

where $\lambda \geq 0$ is a regularization parameter which needs to be tuned a priori. It controls the prior information we have about how close $w_0$ is to 0, and how all weight vectors $w_t$ are close to $w_0$. The first condition is encouraged for high values of $\lambda$ whereas the second condition becomes more important for values of $\lambda$ close to 0. Note that the reference vector $w_0$ corresponds to the commonalities, $C$ in problem (II.1), learned in the transfer stage. Consequently, by applying this model at the calibration stage, we encourage the weight vector of the target subject to be close to this reference vector.

C. Multi-Task Feature Learning

A different way of characterizing how tasks are related is by assuming that they share a low dimensional representation of the data. In other words, one assumes that the tasks’ vectors $w_t$ are linear combinations of a few common basis vectors which need to be estimated from the data. This assumption is studied in [1], where the authors propose the following optimization problem:

$$\min_{U,A} \sum_{t=1}^{T} \left\| X_t^T U a_t - Y_t \right\|^2 + \gamma \|A\|_{2,1}^2 \quad (II.4)$$

s.t.: $A \in \mathbb{R}^{d \times T}$, $U \in \mathbb{R}^{d \times d}$, $U^T U = I$.

That is, the matrix $U$ is used to rotate the data, so that there are some projections which can be useful for all tasks. Consequently, $U$ is constrained to be an orthonormal matrix. The matrix $A$ contains, for each column $t$, the weights of task $t$ for the components learned in $U$. Since the original assumption was that all tasks use the same low dimensional representation of the data, an $\ell_{2,1}$-norm based regularization term is added so that $A$ is encouraged to have only a few non-zero rows.

For a better understanding, we can explore the similarities between this approach and the unsupervised approach Principal Component Analysis (PCA) [18]. The latter procedure obtains a set of components (linear combinations of the attributes) so that they have the largest possible variance whereas in eq. (II.4), the components are obtained so that they are as useful as possible for all tasks. The product $U a_t$ makes the problem (II.4) non convex, yet the authors show that this is equivalent to the following convex problem:

$$\min_{W, D} \sum_{t=1}^{T} \left\| X_t^T w_t - Y_t \right\|^2 + \gamma \sum_{t=1}^{T} w_t^T D^{-1} w_t \quad (II.5)$$

s.t.: $W \in \mathbb{R}^{d \times T}$, $D \in \mathbb{R}^{d \times d}$, $D \succeq 0$, $\text{tr} (D) \leq 1$.

where $\tilde{W} = U A$. Note that eq. (II.5) is an instantiation of problem (II.1) where $\Omega(w, D) = \left\{ \begin{array}{ll} \sum_{t=1}^{T} w_t^T D^{-1} w_t & \text{if } D \succeq 0, \text{tr} (D) \leq 1 \\ \infty & \text{otherwise} \end{array} \right.$.

This approach is studied in [10] and makes the assumption that the divergence among tasks can be expressed in low dimensionality, that is, the tasks are a low rank perturbation of a common “mean task”. It can be seen as a combination of the previous two approaches as the tasks are assumed to be near a reference task $w_0$, as in eq. (II.3), and the differences between the tasks and $w_0$ can be expressed as linear combinations of a few components, as in eq. (II.5).

The equation to minimize is as follows.

$$\min_{W, D} \sum_{t=1}^{T} \left\| X_t^T w_t - Y_t \right\|^2 + \gamma \sum_{t=1}^{T} (w_t - w_0)^T D^{-1} (w_t - w_0) + \lambda \|w_0\|^2$$

s.t.: $W \in \mathbb{R}^{d \times T}$, $D \in \mathbb{R}^{d \times d}$, $D \succeq 0$, $\text{tr} (D) \leq 1$.

It is easy to check that this problem has the form of eq. (II.1), where the commonalities learned in the transfer learning ($C$ in eq. (II.1)) are defined in both elements $D$ and $w_0$ and the regularizer takes the form $\Omega(w, D) = \left\{ \begin{array}{ll} \frac{1}{\gamma} \|w_0\|^2 + \sum_{t=1}^{T} w_t^T D^{-1} w_t & \text{if } D \succeq 0, \text{tr} (D) \leq 1 \\ \infty & \text{otherwise} \end{array} \right.$.

III. DATA DESCRIPTION

Two different datasets were selected to verify the generality of our approach to different modalities and types of tasks: recognition of levels of pain expressions, and recognition of level of muscle activation during physical exercise.

1) UNBC-McMaster Shoulder Pain Expression Archive - This is a publically distributed database (for a detailed description see [15]) which contains facial expressions from 25 subjects who suffer from shoulder pain while they performed active and passive exercises. We seek to predict the level of pain from non acted expressions exhibited from the face. We use all 66 tracked landmark points extracted (by way of Active Appearance Models [15]) from video. From this the distance values of 147 edges were used as the initial feature set, see Figure IV.1. The target outcome is the expressed pain level which was rated using the Prkachin and Solomon Pain Intensity (PSPI) scale; a metric for defining pain based on facial action units.

2) Multi-Modal Chronic Lower Back Pain Dataset - this dataset contains motion capture data and synchronized muscle activity signals acquired using wireless surface Electromyographic (sEMG). From this dataset, we aim to predict
the lumbar paraspinal muscle activity from a range of full body positions. Since this activity has been shown to be an indicator of chronic pain-related behaviour [12, 25], creating robust motion based predictors of sEMG could lead to non intrusive methods to pain in markerless motion capture rehabilitation technology. The sEMG probes were adhered to the skin on the lumbar paraspinal (LPS) muscles. Body movement information was acquired using a Velcro strapped suit containing Inertial Measurement Units (IMU), (Animazoo IGS-180). This suit consisted of 18 IMU devices attached to each body segment capturing motion data at 60Hz. The subjects in this dataset consisted of diagnosed chronic lower pain patients. Each of whom undertook a prescribed set of simple exercises: one leg stand, repeated sit to stand, reaching forwards, bending and a short walk. The data at each time instance is translated to a matrix of triplets indicating the location of 26 anatomical nodes in a global co-ordinate frame. In order to define the body configuration, the relative distances between each of the following nodes are calculated: crown of the head, hands, segmental midpoint of the spine, hip, upper arms, thighs and shanks. The target outcome is the smooth upper boundary of the rectified sEMG signal to characterize the degree of LPS muscle activity. A delay of 100 frames is applied to the motion capture stream since the onset and offset of muscle activity usually precedes the manifest movement.

IV. Results

The three aforementioned TL based approaches: RMTL, MTFL and CMTL are compared with three baseline methods on both datasets. The baseline methods are based on Ridge Regression which is derived from a variant of the minimization expression in eq. (II.1); where the regularization term is replaced by a Frobenius norm. The resultant model can be viewed as a special case of the MTL model without a term to capture commonality information. All experiments have the following settings in common. If \( T \) is the total number of subjects within a dataset, then \( T − 2 \) randomly chosen subjects will form the transfer set. From the remaining two subjects, one will be used for validation and the other as a target subject. The validation subject is used to tune the hyperparameters within the six approaches and the target subject is used to assess their accuracy. From this we follow a leave-one-subject-out validation process (LOSO). Overall whole process is repeated \( R \) times and the corresponding averaged results are shown for each experiment. For each of the transfer subjects, \( m \) instances are sampled for the transfer stage. From the validation and test subject, a set of \( n \) instances are sampled to be used as training set in the calibration stage. In the validation procedure, we considered the values \( 10^k \), with \( k \in \{-6, -4, \ldots, 6, 8\} \) for all hyperparameters.

The three baseline methods and the three TL based methods applied to all experiments are:

- Ridge Regression (RR): In this model the parameters are not trained on any transfer subject information, only the instances of the target subject are used. This is implemented simply to measure the improvement when transfer subjects information is used.
- Grouped Ridge Regression (GRR): The parameters in this model are learned by grouping the training instances of all tasks from all subjects (both transfer and target).
- Grouped Ridge Regression with mean subtracted bias (GRR-b). The method for removing idiosyncratic bias outlined in [4] is applied to the instances prior to the GRR model. The averages to all instances belonging to one subject are subtracted.
- Regularized MTL (RMTL): It applies the approach explained in subsection II.B to perform TL.
- Multi-Task Feature Learning (MTFL): TL is carried out as proposed in subsection II.C.
- Composite Multi-Task Feature Learning (CMTL): It performs TL by applying the approach described in subsection II.D.

A. Pain Level Recognition from Faces

The objective is to build a model able to predict degree of pain from frame by frame facial expression. Since the level of pain is an integer valued scale between 0 - 15 (PSPI score scale), we employ correlation as a performance measure. The absolute distances between facial points are used as the descriptors in each time frame (see Fig. IV.1). We have not used patient 15 since all the instances have a 0 PSPI level. Therefore, \( T = 24 \) subjects. Since the dataset is highly unbalanced in terms of PSPI label distribution (82% of the instances have 0 PSPI level), the training instances for each patient have been selected randomly so that half of them have PSPI > 0.

Fig. IV.1. Face edges used as descriptors of the instances of the UNBC-McMaster Shoulder Pain Expression Archive.

Two experiments are carried out to test the effect of varying \( m \) with a fixed \( n \) and also varying \( n \) with a fixed \( m \). For the first experiment, the number of instances in the training set of the target (and validation) task has been set to \( n = 6 \) and the training set size for the transfer tasks varies in the range \( m = \{10, 30, 50, 70, 90\} \). The process is repeated \( R = 75 \) times and the averaged results are presented in figure IV.2 (top). In the second experiment, we have fixed the training set size of the transfer tasks and vary the number of instances of the target tasks \( n = \{2, 4, 6, 8, 10\} \) in order to study the learning curve. The average results over \( R = 50 \) trials are presented in figure IV.2 (bottom).
B. LPS Muscle Activity Prediction using Motion Capture

Similar to the experiments described in section IV.A, the two $m$ and $n$ variation tests are carried out on the body motion data. For these experiments $T = 6$ subjects. The number of instances in the training set of the target (and validation) task has been set to $n = 15$ and the training set size for the transfer tasks varies in the range $m = \{10, 30, 50, 70, 90\}$. The process is repeated $R = 150$ times and the averaged results are presented in figure IV.3 (top).

For the second experiment we vary $n = \{4, 8, 12, 16, 20\}$. Overall the process is repeated $R = 200$ times and the averaged results can be seen in figure IV.3 (bottom).

V. DISCUSSION AND CONCLUSION

In both comparative experiments described in sections IV.A and B a twofold trend can be seen: firstly among the non TL based methods it can be seen the application of the subtractive bias removal delivers a distinct improvement. The correlation coefficient for GRR-b is consistently higher than the scores for GRR and RR, as can be seen in figures IV.2 (top) and IV.3 (top). Secondly, within the TL based methods RMTL consistently outperforms CMTL and MTFL. Within the comparison between the TL based methods and non TL based methods it can be seen that GRR-b has returned a very similar performance to MTFL and CMTL for muscle activity prediction but is outperformed by RMTL. For PSPI prediction GRR-b is outperformed by all TL based methods. Figure IV.2 (bottom) shows a consistent monotonic improvement given an increase in $n$ for the TL based methods. Figure IV.4 (bottom) repeats this for RMTL and MTFL. With the exception of RR, the non TL based methods show no improvement with the inclusion of a small number of instances $n$ in their training. The improvement in RR is expected since this model is entirely trained on the $n$ instances. A notable difference between the
small increment in the training set size will have little effect. Further instances of training samples. Therefore this relatively procedure where the $n$ instances will only serve as a few further instances of training samples. Therefore this relatively small increment in the training set size will have little effect. In contrast, an increase in $n$ has a far more significant impact on the results produced by the TL-based methods. This demonstrates the power of a few calibration instances in the TL framework and lends itself to modeling idiosyncrasy in an effective way. To test if the results between the various methods were statistically different, we applied a paired $t$-test. The improvement of RMTL was statistically significant with respect to GRR-b, obtaining $p = 1.4 \times 10^{-15}$ for the face results and $p = 2.3 \times 10^{-4}$ for the body movement outputs.

The superiority of GRR-b over GRR and RR supports the findings in [4] showing that idiosyncratic bias removal improves the effectiveness of recognition models. However, it can be seen that RMTL correlates with the ground truth values from both datasets consistently better compared to all other models including GRR-b, see Figure IV.4 for a sample of this from the sEMG predictions. This could be due to the appropriateness of the weight vector assumptions regarding the commonality information in RMTL. Finally it should be noted that both GRR and GRR-b have an advantage compared to TL methods since the former ones are able to access to transfer instances in the calibration stage whereas the latter set of methods only use the learned model from the transfer stage.

References