Almaev, Timur and Martinez, Brais and Valstar, Michel F. (2015) Learning to transfer: transferring latent task structures and its application to person-specific facial action unit detection. In: ICCV15, International Conference on Computer Vision, 11-18 Dec 2015, Santiago, Chile.

Learning to transfer: transferring latent task structures and its application to person-specific facial action unit detection

Timur Almaev Brais Martinez Michel Valstar
The School of Computer Science
University of Nottingham
{psxta4, Brais.Martinez, Michel.Valstar}@nottingham.ac.uk

Abstract

In this article we explore the problem of constructing person-specific models for the detection of facial Action Units (AUs), addressing the problem from the point of view of Transfer Learning and Multi-Task Learning. Our starting point is the fact that some expressions, such as smiles, are very easily elicited, annotated, and automatically detected, while others are much harder to elicit and to annotate. We thus consider a novel problem: all AU models for the target subject are to be learnt using person-specific annotated data for a reference AU (AU12 in our case), and no data or little data regarding the target AU. In order to design such a model, we propose a novel Multi-Task Learning and the associated Transfer Learning framework, in which we consider both relations across subjects and AUs. That is to say, we consider a tensor structure among the tasks. Our approach hinges on learning the latent relations among tasks using one single reference AU, and then transferring these latent relations to other AUs. We show that we are able to effectively make use of the annotated data for AU12 when learning other person-specific AU models, even in the absence of data for the target task. Finally, we show the excellent performance of our method when small amounts of annotated data for the target tasks are made available.

1. Introduction

Automatic facial expression recognition is an active topic in computer vision and machine learning. It has seen so much activity that it already contributed to the creation of three new research directions: Affective Computing [17], Social Signal Processing [24], and Behaviomedics [22]. Imbuing machines with the ability to correctly identify the non-verbal cues humans express, such as facial expressions, would certainly allow a whole new level of interaction between a human being and a machine.

The problem of reliable automatic facial expression recognition is a complex one due to the very high level of variability introduced by factors unrelated to facial expressions, such as identity, head pose or variations in the lighting conditions. The problem is even more complex when we consider non-prototypical facial expressions. There are only six prototypical expressions (often referred to as the six basic emotions; anger, disgust, fear, happiness, sadness and surprise), which makes for a nicely tractable problem from a computer science perspective. Unfortunately, people do not frequently display such strong expressions as anger or disgust in everyday life [24]. Instead, a much wider range of mental states and social intentions are communicated. An often cited number of facial displays shown in day to day life is 7,000.

Figure 1. Example of facial displays and their constituent Action Units (AUs).

To simplify the decoding of such a vast expression space, many researchers take the principled approach of recognising the individual facial muscle actions that contribute to make up a facial expression. Most often the Facial Action Coding System (FACS) [6] is used for this. It was originally developed by Ekman and Friesen in 1978 [7], and revised in 2002 [6]. The revision specifies 32 atomic facial muscle actions, named Action Units (AUs), and 14 additional Action Descriptors (ADs) that account for miscellaneous actions. FACS is comprehensive and objective in its description. Since any facial expression results from the activation

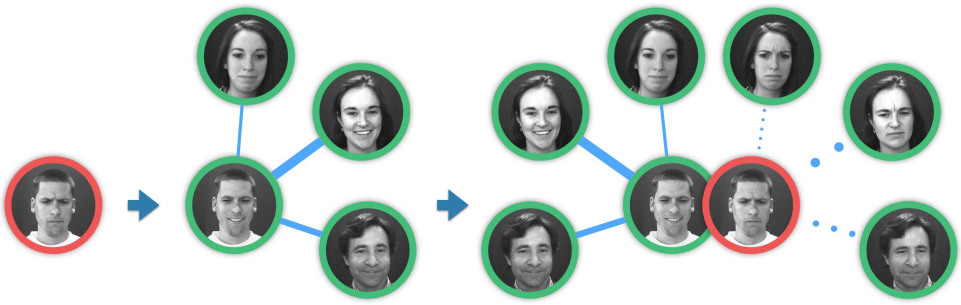

Figure 2. **Overview of Regularised Latent Task Structure** Transfer Learning of task AU12 with known ground truth to un-annotated task AU4. Relations between subjects of learned parameters for AU12 are assumed to be a good initialisation for AU4, and used to constrain its latent structure.

of a set of facial muscles, every possible facial display can be comprehensively described as a combination of AUs [7] (as shown in Fig. 1).

An important issue for the automatic analysis of facial action units remains the poor generalisability to unseen data, in particular to unseen data of new subjects. Even state-of-the-art methods [5, 10, 21] are trained and optimised on relatively small amounts of laboratory-recorded data, limiting their ability to perform adequately under test-time conditions. This problem is exacerbated by the recent interest in applying expression recognition under so-called in-the-wild conditions [15]. That is to say, the community is increasingly veering towards considering unconstrained test-time scenarios, boosting test-time data variability. Inevitably, any serious attempt to cope with this includes acquiring annotated data under these novel conditions. Without automatic help, this effort is largely unapproachable due to both its time-consuming nature and the scarce availability of highly expert annotators. This paper presents a novel transfer learning method that will be of great value for moving towards semi-automatic annotation.

The learning problem becomes simpler if it can be broken down into a set of sub-problems, each of them comprising significantly lower variability in the relevant data. Specific instances of this approach would be the creation of pose-specific or person-specific models. In this work we focus on the latter. Our aim is to train an AU detector tailored to the specific characteristics of a target subject. However, we avoid the need of a full set of person-specific training data. The training of person-specific models can benefit from considering a joint learning problem where all person-specific models are learnt together. In this way the learning process can exploit the commonalities between the tasks and reduce the need for large training sets.

A natural framework to cast this problem in is that of Transfer Learning (TL) and associated frameworks such as Domain Adaptation (DA) and Multi-Task Learning (MTL) [4, 19, 3]. TL techniques are often divided into inductive and transductive [16]. Inductive techniques are those that exploit annotated data both in the source domain (the domain knowledge is transferred from) and in the target domain (the domain knowledge is transferred to). The inductive scenario is often tackled from the perspective of MTL [19, 3]. Instead, for the transductive scenario, TL is tackled from the DA perspective, as no labelled data is available for the target task [4]. MTL and TL have been previously applied to facial expressions recognition and facial AU detection problems. For example, [28] defined each task to relate to a different face region, and used MTL to find a sparse set of face regions capturing facial expressions. Instead, [27] used MTL to detect correlated groups of AU jointly.

In this paper we focus on exploiting the MTL framework for TL. A task is defined as a binary classification problem for a specific AU and a specific subject. Tasks are in this case related to others tasks when the subject varies but the AU is the same, or when the subject is the same but the AU varies. Defining the problem in these terms results in a tensorial structure. This duplicity of relations regarding AU and subjects has already been noted and exploited in the literature [18].

Similar to [18], our work also exploits the tensorial relations between tasks. However, we our approach differs both on the scenario considered and in the technical approach followed. We assume an asymmetry on the annotated data, as we consider one AU to be a reference task. Sufficient annotated data is available for the reference task, including for the target subject. In practice this will be an AU that is comparatively easy to elicit and to annotate, such as a smile (AU12). However, the amount of data available for other AUs can be limited or even non-existent.

This scenario is justified by the practicalities of data acquisition: eliciting and annotating a smile is very easy. Instead, capturing expressive information for e.g. sadness, anger, or less common or subtle AU, can be very challenging. This situation is not only common at test time, but also typical for currently existing datasets, where subjects sel-

dom exhibit the exhaustive range of AU targeted and data is much more frequent for some AU than for others. We summarise this scenario in the research question we are considering: *Can we make use of annotated data for a specific facial expression/AU, and transfer this information to build subject-specific models of other facial expressions/AUs with no or very little annotated data?*

Our approach to answer this question is as follows: inspired by the GOMTL approach [11], we develop a Transfer Learning approach by first considering a learning problem for each AU where the structure of task relations is estimated. We then harness the tensorial information by constraining *the underlying latent relational information*. Our reasoning is that while the optimal parameters for different subjects will be similar, this is not true when varying the AU instead. For example, optimal parameters for AU12 and AU1 are unlikely to be close in the parameter space even for a fixed subject identity, as AU12 is mouth-related and AU1 is eye-related. However, this is different when we consider the latent relations among subjects. The fact that subjects are related for a specific AU is likely to be based on shared appearance characteristics (gender, age, having a beard, etc). These relations are thus likely to be valid for other AUs. Thus, our aim is to capture the latent relations between subjects using a specific easy-to-annotate, easy-to-elicit and easy-to-detect AU, and then transfer this information to the problem of learning to detect other AUs.

In summary, the main contributions of our work are:

- We define a new TL scenario for automatic AU analysis reflecting the practicalities of data acquisition.
- We propose an extension of the MTL framework capable of harnessing latent structures in problems with tensorial structure.
- We show the effectiveness of our TL approach even in the extreme case where no annotated data is available for the target task, obtaining better prediction accuracy than any other existing MTL-based TL method.
- We show that adding small amounts of labelled data of the target task very quickly improves performance, staying above any other method for any quantity of annotated data on both datasets tested.

2. Literature Review

MTL and TL techniques: MTL approaches can be grouped into techniques that regularise the parameter space, and techniques that learn the relevant features jointly. Multi-task feature learning tries to avoid spurious correlations between the features and the labels by learning jointly which features are important for inference. Examples of this approach are [1] and [18], where all the tasks are learnt on a shared sparse subset of the original features.

MTL techniques that regularise the parameter space assume instead that related tasks result in related optimal parameters. Similarities in the parameter space can be harnessed through either a soft constraint (e.g. being close in an Euclidean sense), or a hard constraint (e.g. lying on a subspace). A notable example is that of [9], where a regulariser was used to enforce the task parameters to be close to the mean of all task parameters in an L_2 sense. This work was extended in [8], where binary relations among tasks could be defined manually. A common setting of this framework is to use the average of pairwise distances between task parameters as a regulariser. At its core, these methods use the optimal parameters for all the tasks to build an empirical regulariser over the parameter space.

These methods assume that all of the tasks are related in the same way. This can lead to sub-optimal performance or even to the so-called negative transfer, i.e., the pernicious effect introduced by learning unrelated tasks together. This observation has led to recent works exploring different ways of defining the relations in a more flexible manner. Recently, some works have aimed at automatically estimating the structure of relations among the different tasks [11, 13, 29]. That is to say, these works find a latent structure that reflects the relations among tasks, allowing for selective transfer.

We pay special attention to [11] and [13]. Both of these works regularise the parameter space by constraining all of the task parameters to lie in a shared subspace. Furthermore, the subspace is learnt by making use of sparse coding techniques, so that the target tasks are explained using only a handful of the dimensions of the subspace. This approach has a strong relation with sparse coding for face recognition [25], and relies on the concept that two examples are only related if they contribute to explain each other as succinctly as possible. Learning within this framework proceeds by alternating the learning of a set of basic tasks (the generator of the subspace), and learning the parameters for each of the tasks, expressed now in terms of the sparse coefficients within the linear subspace.

To the best of our knowledge, the only existing attempt to harness task relations within a tensorial structure is that of [18]. However, the aim and technical approach is very different from ours. Specifically, 1) [18] corresponds to the feature learning MTL family, while our work belongs to the parameter regularisation family. No effort has been done so far to harness these relations from the perspective of parameter constraints 2) unlike [18], we account for different levels of relatedness among tasks 3) our method explicitly considers the case of TL while [18] does not. It is interesting to note that the tensorial structure stems naturally from the data rather than because of a particularity on the system design. Furthermore, since AU occur at different parts of the face and relate to different facial actions, it is counter-

intuitive to think that optimal parameters across AU should "look alike". The core aspect of our approach is realising that parameter across AU should not be close, but rather they should have a similar latent structure of task relations. This is a profound change in perspective and it corresponds to an intuitive and natural yet powerful aspect of the nature of the data.

Person-specific models for automatic AU analysis: The creation of person-specific models using TL techniques has only very recently been addressed. Of the works doing so, some have aimed at transductive TL (i.e., TL without making use of labels for the target subject). For example, [4] proposed a re-weighting of the source (training) distribution to match that of the domain (test) distribution. The classifier is then modified to be optimal with respect to the weighted training set. A similar approach, also relying on the weighting of the training examples to minimise the distribution mismatch, was proposed in [3]. A different idea was followed in [20] and in [26], where the authors proposed to learn discriminative mappings from the space of training distributions to the parameter space. To this end, they trained a set of person-specific models, used as the training examples to learn the mapping. A kernel representation to measure similarity between distributions was employed.

On the inductive TL side, some works have considered the creation of person-specific models in the presence of annotations for the target domain. For example, [3] also proposed a method for the creation of person-specific models based on a boosting TL technique. A set of basis classifiers were computed in the source domains, and then linearly combined in the target domain. By employing both a Transductive and an Inductive TL technique they were able to objectively measure the gain of using labelled data of the target task.

Instead, [19] explored different formulations of MTL for the problem of person-specific AU detection, comparing a parameter space constrained method [8] and two MTL feature learning approaches. Finally, [18] presented a multilinear extension to the MTL convex feature learning approach of [1]. The learning of person-specific AU models is one of the applications chosen for the experimental validation of their MTL framework due precisely to the tensorial nature of the AU-subject relations.

Our approach has an inherently different aim from Transductive TL approaches, as we assume some amount of easily obtained labelled data is available. We distinguish ourselves from previous Inductive TL approaches in that the different AU play an asymmetric role, and in that it is our aim to exploit the tensorial relations between tasks.

3. Learning the latent task relations

In this section we first review the work presented in [11]. This methodology is used for finding the latent relations among tasks when organised in matrix form. That is to say, we first consider the problem of creating subject-specific models for one specific AU (independently of other AUs), and review the methodology used to estimate the underlying structure of tasks relations. In Sec. 3.2 we extend this technique by incorporating information from the tensorial structure so that we consider two modes of relations between tasks: subjects and AU. We will do so by relating and constraining the latent structure of relations learnt for each of the different AU-specific MTL problems. The resulting extended problem is then minimised jointly through alternated gradient descent.

3.1. Finding Latent Relations Among Subjects

Grouping and Overlap in Multi-Task Learning (GOMTL) [11] aims to improve classification performance for a number of related tasks by learning them jointly, simultaneously discovering the degree of mutual relationship between tasks and exploiting these relations for learning the individual task parameters. Let T be the number of tasks[1] and $\mathbf{Z}_t = \{(\mathbf{x}_t^i, y_t^i)\}_{i=1,\dots,N_t}$ be the training set for task t. The goal is to learn the parameter matrix $\mathbf{W}$ of size $d \times T$, where d is the feature dimensionality and T the number of tasks. By $\mathbf{W}_{:,t}$ we indicate the column t of matrix $\mathbf{W}$, which stores the parameters of task t.

The idea of GOMTL is to constrain the parameter space by imposing that all the task parameters must lie on a common linear subspace. It is thus assumed that there are K basis tasks that are the generators of this subspace, and every observed task is then represented as a linear combination of the basis tasks. This assumption makes it possible to write matrix $\mathbf{W}$ as:

$$\mathbf{W} = \mathbf{L}\mathbf{S} \tag{1}$$

where $\mathbf{L}$ contains the parameters of the basis tasks, resulting in a $d \times K$ dimensionality, and $\mathbf{S}$ is the $K \times T$ matrix containing the linear coefficients for the tasks. In order to favour grouping of tasks, a sparsity constraint is imposed on the linear coefficients of each task. The resulting loss function then takes the following general form:

$$\mathcal{E}(\mathbf{L}, \mathbf{S}) + \lambda \|\mathbf{S}\|_1 + \mu \|\mathbf{L}\|_F^2 \tag{2}$$

where the first term is defined as:

$$\mathcal{E}(\mathbf{L}, \mathbf{S}) = \sum_{t=1}^{T} \sum_{i=1}^{N_t} \mathcal{L}\left(y_t^i, \mathbf{L}'\mathbf{S}'_{:,t}\mathbf{x}_t^i\right) \tag{3}$$

[1] Bold lower-case letters indicate (column) vectors. Matrices are indicated with upper-case bold typeset letters. All non-bold letters are scalars.

That is to say, $\mathcal{E}$ is the accumulated empirical error term of all tasks, the ℓ_1 regulariser imposes independent sparsity constraints over the coefficients of each task, and the typical ℓ_2 regularisation is imposed over each of the K latent tasks. The interaction between the different tasks however comes from the fact that all $\mathbf{W}_{:,t}$ depend on the shared variable $\mathbf{L}$. Through this formulation, the level of relation between tasks is captured in the commonalities of the column-wise sparse parameters of matrix $\mathbf{S}$.

The above loss function is not convex overall. However, it is convex in $\mathbf{L}$ for a fixed $\mathbf{S}$ and vice-versa. In consequence, [11] adopted an alternating optimisation strategy, first minimising for $\mathbf{S}$ with a fixed $\mathbf{L}$, and then minimising for $\mathbf{S}$ while fixing $\mathbf{L}$. More formally, we first solve T independent minimisation problems:

$$\mathbf{S}_{:,t} = \underset{\mathbf{s}}{\operatorname{argmin}} \sum_{i=1}^{N_t} \mathcal{L}(y_t^i, \mathbf{s}'\mathbf{L}'\mathbf{x}_t^i) + \lambda\|\mathbf{s}\|_1, \qquad (4)$$

followed by the minimisation of:

$$\underset{\mathbf{L}}{\operatorname{argmin}}\ \mathcal{E}(\mathbf{L}, \mathbf{S}) + \mu\|\mathbf{L}\|_F^2. \qquad (5)$$

This alternating minimisation procedure is initialised by training T independent models, storing them in a matrix $\mathbf{W}^{(0)}$, and then computing an SVD decomposition of $\mathbf{W}^{(0)}$. $\mathbf{L}$ is defined as the set of eigenvectors corresponding to the K largest eigenvalues.

It is interesting to note that no form for the error term has been defined yet. This highlights the flexibility and generality of this formulation. Since we are addressing a binary classification problem, in our experiments we use a Logistic Regression loss function.

While this algorithm has been shown to outperform single task as well as a number of MTL approaches, it fails however to harness and exploit tensorial relations. It is thus necessary for its practical application to AU problems to have a manually annotated set of examples *for each AU*, resulting in an unrealistic scenario.

3.2. Regularising the Latent Relations

Let us now extend the notation to allow for two modes of variation within the tasks. Specifically, a task will now be indexed by subject, $t_1 \in \{1, \ldots, T_s\}$, and AU, $t_2 \in \{1, \ldots, T_{AU}\}$. Let $\mathbf{Z}_{t_1,t_2}$ denote the per-task training set. $\mathbf{W}$ is now a tensor of dimensions $d \times T_s \times T_{AU}$. The parameters of task $\{t_1, t_2\}$ is now noted $\mathbf{W}_{:,t1,t2}$. The same notation holds for $\mathbf{S}$ and $\mathbf{L}$.

We first consider, for each AU, the learning problem as defined in Sec. 3.1. That is to say, we consider a GOMTL problem for each AU $\in \{1, \ldots, T_{AU}\}$. This consists of learning a matrix of weights $\mathbf{W}_{:,:,t_2}$ so that it is decomposed into $\mathbf{L}_{:,:,t_2}$ and $\mathbf{S}_{:,:,t_2}$. We however extend the loss resulting from combining all these problems. Our extended loss function is defined as:

$$\sum_{t_2=1}^{T_{AU}} \mathcal{E}(\mathbf{L}_{:,:,t_2}, \mathbf{S}_{:,:,t_2}, \mathbf{Z}_{:,t_2}) + \mu\|\mathbf{L}\|_F^2 + \lambda\|\mathbf{S}\|_1 + \mathcal{R}(\mathbf{S}) \quad (6)$$

The relations between the T_{AU} GOMTL models is harnessed through $\mathcal{R}(\mathbf{S})$. Let us first explain the intuition behind the proposed regulariser. Firstly, we note that all the tasks for a given AU can reasonably be expected to be related, in the sense that their optimal parameters should be close. Instead, this property does not hold for tasks across AU, mainly because different AU are localised in different parts of the face and thus the optimal parameters will not be close[2]. Directly regularising the task parameters across AUs would thus result in a case of negative transfer. However, we note that instead the **latent relations** can be constrained. That is to say, if subject i is related to subject j for a specific AU, then both subjects are likely to be related for any other AU. We capture this intuitive idea by defining a regulariser over the latent structures across different AUs.

It is possible now to apply the same principles that have been used previously for parameter regularisation within MTL, such as the mean-regularised MTL [9], the pairwise regularisation of [8], or even to apply again the same idea of GOMTL on $\mathbf{S}$. However, we invoke now our scenario of interest: we consider that some AUs are much easier to elicit, annotate, and even detect automatically than others. Of those, AU12 (a smile) is probably the most paradigmatic case, as smiles are easily identifiable (see for example the results on the FERA challenge [23]). We then refer to AU12 as the reference AU. Let us thus re-define the notation to highlight the asymmetry of the role between tasks. Let $t_* \in \{1, 2, \ldots, T_{AU}\}$ denote the reference AU, for which we assume that the associated training sets $\mathbf{Z}_{t_1,t_*}$, $t_1 = 1, \ldots, T_s$ contain sufficient training examples, including for that of the target subject. What we aim is to capture the latent structure of relations between subjects using AU12, and then transfer the latent structure to any other AU. Then we define the regulariser over the latent structures as:

$$\mathcal{R}(\mathbf{S}) = \tau \sum_{t=1, t\neq t_*}^{T_{AU}} \|\mathbf{S}_{:,:,t} - \mathbf{S}_{:,:,t_*}\|_F^2 \qquad (7)$$

The minimisation relies again on alternating minimisations. Specifically, it is possible to loop over the tasks, first minimising:

$$\mathbf{S}_{:,t_1,t_2} = \underset{\mathbf{s}}{\operatorname{argmin}}\ \mathcal{E}(\mathbf{L}_{:,t_1,t_2}, \mathbf{s}, \mathbf{Z}_{t_1,t_2}) + \lambda\|\mathbf{s}\|_1 + \tau\|\mathbf{s} - \mathbf{S}_{:,t_1,t_*}\|_2^2 \qquad (8)$$

[2] It is actually common to use a different set of features, e.g., upper face features for upper face AU

where the last term vanishes if $t_2 = t_*$. Then we proceed by minimising $\mathbf{L}_{:,:,t_2}$ looping over t_2 in an identical fashion to that in Eq. (5).

Let us now consider the Transfer Learning scenario explicitly. We assume that there exists a reference task t_*, for which all subjects have annotated data. Let us simplify this scenario by considering only one target task at a time. That is to say, we consider only one AU at a time besides the reference AU, and we aim to learn a model for that AU for a specific subject n making use of very few or even no annotated data of the target task. In the latter case (the most interesting in terms of applicability), the constraint imposed by Eq. (7) means that the latent structure will be transferred directly, i.e., $\mathbf{S}_{:,t_n,t_2} = \mathbf{S}_{:,t_n,t_*}$, while the latent tasks $L_{:,:,t_2}$ and $L_{:,:,t_*}$ remain the same.

In fact, we can understand the regularisation in Eq. (7) as an extreme case of an empirical prior over the latent structure $\mathbf{S}_{:,t_n,t_2}$. It is perfectly feasible to consider more than one reference task, and in this case the interpretation as an empirical prior would be more natural. However, this would push us away from our scenario of interest. In the presence of annotated data for the target task, the transfer is attained by minimising the joint loss function defined in Eq. (6) by alternating between Eqs. (5) and (8).

The RLTS learning process is described in algorithm 1.

Input:
Z_{t_1,t_2} : Training set for all subjects t_1 and AUs t_2
λ, μ, τ : Regularisation parameters
t_* : Reference AU index
K : Number of latent tasks
Output: Linear predictors $\mathbf{W}$ for all T_{AU} and T_s tasks.
1: Learn all tasks independently to obtain in $\mathbf{W}^{(0)}_{:,:,t_2}$.
2: Initialise $\mathbf{L}_{:,:,t_2}$ for all t_2 as indicated in section 3.1.
while *not converged* **do**
  3: Solve Eq. 8 for all subjects and AU to update $\mathbf{S}$
  4: For all AU, fix $\mathbf{S}_{:,:,t_2}$ and update $\mathbf{L}_{:,:,t_2}$ (Eq. 5)
end
5: Obtain $\mathbf{W}_{:,:,t_2} = \mathbf{L}_{:,:,t_2}\mathbf{S}_{:,:,t_2}$ for all t_2.

Algorithm 1: RLTS - Reguralised Latent Task Structure.

4. Experiments & Results

Data: We have used the DISFA dataset [14] and the UNBC-McMaster Shoulder Pain Expression dataset [12] to perform our experiments. The facial expressions displayed in both datasets are spontaneous. The head is usually kept in a near-frontal pose with respect to the camera. DISFA is annotated for 12 AU out of the possible 32 AU, while the McMaster dataset is annotated for 10 AU. Both databases also provide very accurate landmark locations on a frame-by-frame basis for a total of 66 facial points, which were annotated by the authors in a semi-automatic manner.

AUs	# Subjects	# Positives	# Episodes
1	17	8524	144
2	14	7041	89
4	24	24502	226
5	8	2201	68
6	25	19469	167
9	17	6774	62
12	27	30794	247
15	17	7684	84
17	22	12764	260
20	13	3904	72
25	27	46052	289
26	27	24976	313

Table 1. Action Units statistics on the DISFA dataset. The *subjects* column contains the number of subjects which had enough positives (around 250 per task).

Table 1 shows some statistics regarding the AU occurrence on the DISFA dataset. The table clearly shows how both the number of annotated frames and the number of episodes varies greatly between AUs. Frames within an episode tend to be more correlated. Thus, the number of episodes is the better indicator of the variability of the data, and can also be used as an indicator of how easy it is for a certain AU to be elicited.

Features: We employ a set of geometric features derived from the facial landmark locations. We use the set of 49 inner-facial landmarks and discard the contour landmarks. We then select the set of facial landmarks for which their location does not change with facial expression activation and refer to them as the *stable points*. This set consists in our case of the four eye corners and the nose region.

Each face shape is aligned first to the mean shape of the dataset[3] through a non-reflective affine transformation aligning the stable points of the current frame and the reference shape. The first set of 98 features are simply the difference between the registered shape and the reference shape. The next 98 features are computed as the displacement of the registered shape locations from the previous frame to the current frame. We generate another 49 features by calculating the median of the stable points and computing the Euclidean distance from it to each of the landmarks. The remaining features are extracted from three face regions, the left eyebrow and eye, the right eyebrow and eye, and the mouth region. For each of these regions, features are obtained by taking the Euclidean distance and angle between two pairs of points belonging to the same components.

Task definition: So far there has been no explicit definition

[3] Any other frontal-looking face shape can in any case be used instead as a reference shape.

AUs	SVM	MLMTL	GOMTL	RLTS
1	0.346	0.038	**0.547**	0.541
2	0.516	0.500	0.544	**0.717**
4	0.461	0.460	0.552	**0.588**
5	0.192	0.137	0.176	**0.265**
6	0.708	**0.720**	0.576	0.619
9	0.289	0.287	**0.377**	0.375
15	0.293	0.166	**0.394**	0.376
17	**0.381**	0.319	0.379	0.345
20	**0.336**	0.294	0.199	0.254
25	0.699	0.685	**0.772**	0.740
26	0.699	0.549	0.704	**0.717**
Mean	0.447	0.378	0.474	**0.503**

Table 2. Evaluation results measured in accumulated F1 score when using no training examples of the target task (DISFA).

AUs	SVM	MLMTL	GOMTL	RLTS
1	0.354	0.779	0.719	**0.809**
2	0.620	0.839	0.713	**0.872**
4	0.484	0.872	0.765	**0.874**
5	0.290	0.464	0.274	**0.476**
6	0.736	0.806	0.797	**0.831**
9	0.311	0.728	0.602	**0.770**
15	0.294	0.661	0.575	**0.705**
17	0.437	**0.701**	0.518	0.694
20	0.457	0.559	0.338	**0.628**
25	0.725	0.832	0.824	**0.854**
26	0.724	**0.843**	0.7084	0.699
Mean	0.494	0.735	0.621	**0.746**

Table 3. Evaluation results measured in accumulated F1 score with 60 training instances of the target task (DISFA).

of the empirical error term. The presented framework can be used with any definition of empirical error that is convex and can be minimised through an efficient gradient descent. In our case, we consider a binary problem per AU, and we have used a logistic regression function. The methodology could be readily extended to AU intensity estimation by considering, e.g. a linear regression loss function.

Each task is defined to contain at least 150 positive examples, so that there are enough instances to perform cross-validation. As a consequence, we used only 6 out of the 10 AU annotated in the McMaster dataset, while we used every AU annotated in the DISFA dataset, but we restrict ourselves to a subset of the subjects. For example, there are 27 subjects on the DISFA dataset with enough AU12 annotations, which can result in a slow minimisation procedure when performing cross-validation. When comparing against other MTL methodologies, the tasks for each of the methods are defined over exactly the same training data. That is to say, the partitions are pre-computed and then passed to the learning and testing routines of each method.

Optimisation procedure: GOMTL is initialised as explained in Sec. 3.1, i.e., we run linear SVM with fixed margin parameter $C = 1$ to create independent tasks from which to initialise $\mathbf{L}$. For our method, we proceed by initialising each matrix $\mathbf{L}_{:,:,t_2}$ in this same manner, and then alternate the minimisation of $\mathbf{S}_{:,t_2}$, for all tasks, and $\mathbf{L}_{:,:,t_2}$, for all tasks. In both cases, the minimisation is performed by gradient descent using a line search algorithm. Parameter optimisation was performed using a 5-fold cross-validation strategy. For our method, we optimised the three regulariser parameters (λ in Eq. 6, μ and τ in Eq. 8) and the number of basis tasks K. For MLMTL we optimised all parameters possible, for linear SVM we optimised the margin, and for GOMTL we optimised λ, μ and K as defined in Sec. 3.1. The parameter search was conducted using a simple grid search within a pre-define range of values. If the optimal value was on an extreme of the range, the search was extended.

Baseline methods: We benchmark the performance of our method against linear SVM (for which we use the LIBSVM implementation [2]), GOMTL [11] and MLMTL [18]. SVM is an exemplar of a learning method where the person-specific models are trained independently, and it serves the purpose of highlighting the performance increase we obtain through the use of MTL methods. GOMTL is the most related approach to ours as it captures latent relations among subjects. However, it does not incorporate tensorial task relations. It thus serve the purpose of showing the performance gain we obtain by considering also the task relations across AU. As with our method, GOMTL allows for the use of any error function. We use logistic regression to further improve the relevance of the comparison. Finally, MLMTL is the only MTL besides ours that considers both relations across subjects and across AU. We use the non-convex formulation, reported to be the best in [18].

Evaluation protocol: Throughout our experiments we employ a Leave-One-Task-Out (LOTO) evaluation approach. For the sake of simplicity, we only consider one AU at a time besides the reference task. Performance is reported in terms of the combined F1 error across the whole data set, i.e., the predictions obtained for the different subjects are concatenated into a single vector from which the F1 error is computed (where of course only the predictions for the target subject are obtained for each LOTO step). In this way we correct for composition unbalance on the test set.

Results: We perform an experiment in which we measure performance while increasing the amount of annotated data available for the target task. The results for our method and the baseline methods are shown in Fig. 3 for the DISFA dataset, and in Fig. 4 for the McMaster dataset. We can clearly see how the proposed RLTS method stays atop all of the baseline methods for all amounts of data and for both

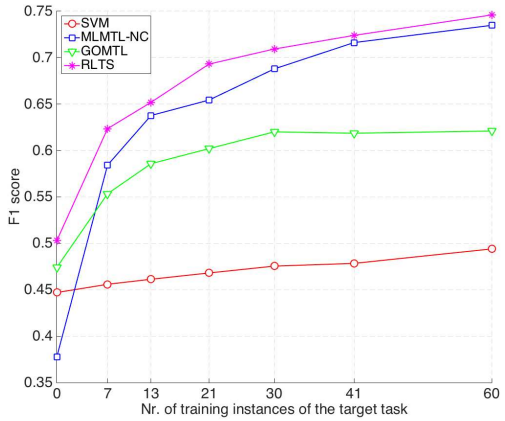

Figure 3. Performance on the DISFA dataset. The y-axis shows the accumulated F1-score, while the x-axis shows the number of examples of the target task used.

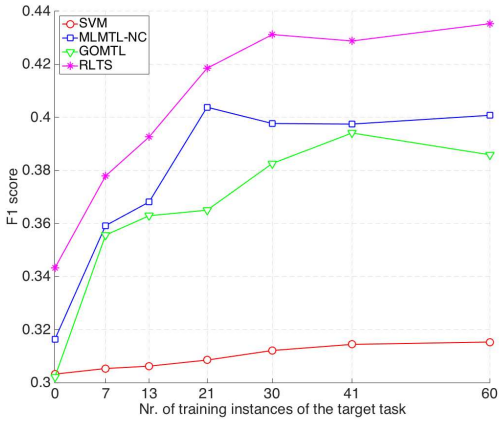

Figure 4. Performance on the McMaster dataset. The y-axis shows the accumulated F1-score, while the x-axis shows the number of examples of the target task used.

datasets.

As a specific case of particular interest, Table 2 summarises the performance on the DISFA dataset of the different algorithms when no training data for the target task is available. While this is a transductive scenario, we can still use the regularisation obtained from the structure for AU12 to contribute to the prediction of the target task. In this case we use the mean task across subjects (other than the target subject) for the evaluation of each of the baseline methods. This is the best guess possible for the case of independent tasks and for GOMTL. While MLMTL also uses tensor information, the problem aims to learn features jointly, and no constraint is imposed on the parameter space. It is thus again only possible in this case to use the mean task. Instead, we are able to do better than using the mean task. This is because our method uses the latent structure learnt from the reference AU, and then applies this latent structure to the target task. Since the latent structure learnt changes for every subject, the resulting parameters for the target task are different for every subject despite having no training data for them. This constitutes one of the major results of our work, as we are learning an empirical prior over the transfer learning process. That is to say, we effectively learn to transfer.

The per-AU performance for the DISFA dataset when using 60 training instances of the target task is shown in Table 3. Remarkably, we obtain a 20% relative average performance increase with respect to GOMTL, highlighting the importance of taking the tensorial structure into consideration. In fact, our method outperforms any other baseline method for all AU except for AU26 and AU17, where performance is marginally smaller than for MLTML.

5. Conclusions & Future Work

In this paper we have introduced a novel MTL and TL approach, called Regularised Latent Task Structure. The experiments show the advantage of the proposed approach over the most relevant state-of-the-art MTL approaches for learning person-specific facial AU detection models. Remarkably, we are able to produce subject-specific AU detection models even without any training data for the target task by exploiting annotated data of the same subject but for a different AU. This allows learning person-specific models for facial expressions only using data easy to elicit, annotate and automatically detect.

While the methodology presented in this work is aimed at the creation of person-specific AU detection models, the framework is naturally described without making any assumption on the loss function definition, except that the error term is convex and smooth. We thus could naturally apply this framework to AU intensity estimation. Furthermore, we assume a tensorial structure on the data. While we consider here different AUs and subjects as factors, this type of relations occurs in many types of data. For example, head pose-specific models are similarly a very natural target. Totally different problems, such as recommender systems, can also be considered: Are two persons sharing their films interests more likely to share their music preferences?

6. Acknowledgements

This work was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement №645378. We are also very grateful for the access to the University of Nottingham High Performance Computing Facility.

References

[1] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008. 3, 4

[2] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Trans. on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. 7

[3] J. Chen, X. Liu, P. Tu, and A. Aragones. Learning person-specific models for facial expression and action unit recognition. *Pattern Recognition Letters*, 34(15):1964 – 1970, 2013. 2, 4

[4] W.-S. Chu, F. De la Torre, and J. F. Cohn. Selective transfer machine for personalized facial action unit detection. In *Computer Vision and Pattern Recognition*, pages 3515–3522, 2013. 2, 4

[5] X. Ding, W. Chu, F. D. la Torre, J. F. Cohn, and Q. Wang. Facial action unit event detection by cascade of tasks. In *Int'l Conf. Computer Vision*, pages 2400–2407, 2013. 2

[6] P. Ekman, W. Friesen, and J. Hager. *Facial Action Coding System (FACS): Manual*. A Human Face, 2002. 1

[7] P. Ekman and W. V. Friesen. *Facial Action Coding System: A technique for the measurement of facial movement*. Consulting Psychologists Press, 1978. 1, 2

[8] T. Evgeniou, C. A. Micchelli, and M. Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:615–637, 2005. 3, 4, 5

[9] T. Evgeniou and M. Pontil. Regularized multi–task learning. In *Int'l Conf. on Knowledge Discovery and Data Mining*, pages 109–117, 2004. 3, 5

[10] B. Jiang, M. Valstar, B. Martinez, and M. Pantic. A dynamic appearance descriptor approach to facial actions temporal modeling. *Trans. on Cybernetics*, 44(2):161–174, 2014. 2

[11] A. Kumar and H. Daumé III. Learning task grouping and overlap in multi-task learning. In *Int'l Conf. on Machine Learning*, 2012. 3, 4, 5, 7

[12] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews. Painful data: The UNBC-McMaster shoulder pain expression archive database. In *Automatic Face and Gesture Recognition*, 2011. 6

[13] A. Maurer, M. Pontil, and B. Romera-Paredes. Sparse coding for multitask and transfer learning. In *Int'l Conf. on Machine Learning*, pages 343–351, 2013. 3

[14] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. Disfa: A spontaneous facial action intensity database. *Trans. on Affective Computing*, 4(2):151–160, 2013. 6

[15] D. McDuff, R. El Kaliouby, T. Senechal, M. Amr, J. F. Cohn, and R. Picard. Affectiva-mit facial expression dataset (AM-FED): Naturalistic and spontaneous facial expressions collected in-the-wild. In *Comp. Vision and Pattern Recog. - Workshop*, 2013. 2

[16] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Trans. on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. 2

[17] R. W. Picard. *Affective computing*. MIT press, 2000. 1

[18] B. Romera-Paredes, H. Aung, N. Bianchi-Berthouze, and M. Pontil. Multilinear multitask learning. In *Int'l Conf. on Machine Learning*, pages 1444–1452, 2013. 2, 3, 4, 7

[19] B. Romera-Paredes, M. S. H. Aung, M. Pontil, N. Bianchi-Berthouze, A. C. de C. Williams, and P. Watson. Transfer learning to account for idiosyncrasy in face and body expressions. In *Automatic Face and Gesture Recognition*, 2013. 2, 4

[20] E. Sangineto, G. Zen, E. Ricci, and N. Sebe. We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer. In *Int'l Conf. Multimedia*, pages 357–366, 2014. 4

[21] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost. Facial action recognition combining heterogeneous features via multi-kernel learning. *Trans. on Systems, Man and Cybernetics, Part B*, 42(4):993–1005, 2012. 2

[22] M. Valstar. Automatic behaviour understanding in medicine. *Proceedings ACM Int'l Conf. Multimodal Interaction*, 2014. 1

[23] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. R. Scherer. Meta-analysis of the first facial expression recognition challenge. *Trans. on Systems, Man and Cybernetics, Part B*, 42(4):966–979, 2012. 5

[24] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759, 2009. 1

[25] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *Trans. on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009. 3

[26] G. Zen, E. Sangineto, E. Ricci, and N. Sebe. Unsupervised domain adaptation for personalized facial emotion recognition. In *Int'l Conf. on Multimodal Interaction*, pages 128–135, 2014. 4

[27] X. Zhang and M. Mahoor. Simultaneous detection of multiple facial action units via hierarchical task structure learning. In *Int'l Conf. on Pattern Recognition*, pages 1863–1868, 2014. 2

[28] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. Metaxas. Learning active facial patches for expression analysis. In *Computer Vision and Pattern Recognition*, 2012. 2

[29] J. Zhou, J. Chen, and J. Ye. Clustered multi-task learning via alternating structure optimization. In *Advances in Neural Information Processing Systems*, pages 702–710, 2011. 3