From Emotions to Action Units with Hidden and Semi-Hidden-Task Learning

Adria Ruiz
Universitat Pompeu Fabra
Barcelona
adria.ruiz@upf.es

Joost Van de Weijer
Centre de Visio per Computador
Barcelona
joost@cvc.uab.es

Xavier Binefa
Universitat Pompeu Fabra
Barcelona
xavier.binefa@upf.es

Abstract

Limited annotated training data is a challenging problem in Action Unit recognition. In this paper, we investigate how the use of large databases labelled according to the 6 universal facial expressions can increase the generalization ability of Action Unit classifiers. For this purpose, we propose a novel learning framework: Hidden-Task Learning. HTL aims to learn a set of Hidden-Tasks (Action Units) for which samples are not available but, in contrast, training data is easier to obtain from a set of related Visible-Tasks (Facial Expressions). To that end, HTL is able to exploit prior knowledge about the relation between Hidden and Visible-Tasks. In our case, we base this prior knowledge on empirical psychological studies providing statistical correlations between Action Units and universal facial expressions. Additionally, we extend HTL to Semi-Hidden Task Learning (SHTL) assuming that Action Unit training samples are also provided. Performing exhaustive experiments over four different datasets, we show that HTL and SHTL improve the generalization ability of AU classifiers by training them with additional facial expression data. Additionally, we show that SHTL achieves competitive performance compared with state-of-the-art Transductive Learning approaches which face the problem of limited training data by using unlabelled test samples during training.

1. Introduction

During years, automatic facial behavior analysis has focused on the recognition of universal facial expressions or Action Units. These two problems are motivated by the well-known studies of the psychologist Paul Ekman. Ekman showed that there exist 6 universal emotions (anger, happiness, fear, surprise, sadness, and disgust) and that each of them has a corresponding prototypical facial expression [7]. Despite their cross-cultural universality, it has been demonstrated that people can perform many other non-basic expressions representing contempt, embarrassment or concentration and that combinations of these expressions are

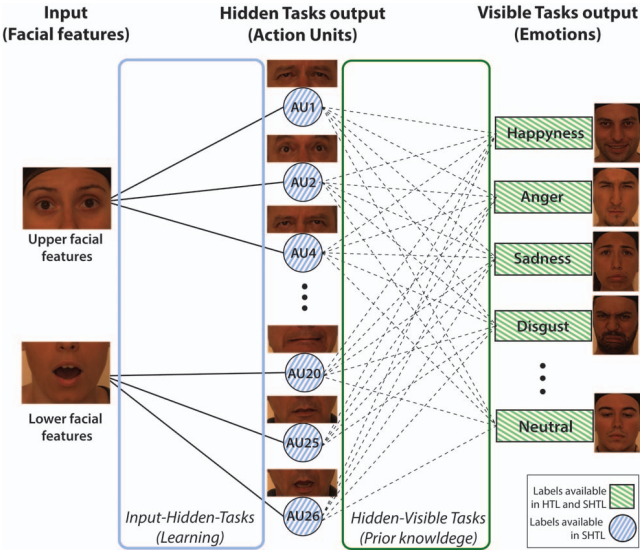

Figure 1: Hidden-Task Learning and Semi-Hidden-Task Learning frameworks applied to Action Unit recognition. HTL aims to learn AU classifiers (Hidden-Tasks) by using only training samples labelled with universal facial expressions (Visible-Tasks). For this purpose, HTL exploits prior knowledge about the relation between Hidden and Visible-Task outputs. In this work, the relation between Action Unit and facial expressions is modelled based on empirical results obtained in psychological studies. SHTL is an extension of HTL assuming that samples from the Hidden-Tasks (Action Units) can also be provided. We show that the use of additional facial expression training samples increases the generalization ability of the learned AU classifiers.

usual in our every-day life [5]. For these reasons, a more objective method to categorize expressions is by using the Facial Action Coding System (FACS) [6]. In FACS, Ekman defined a set of 45 Action Units which are atomic movements in the face caused by the activation of one or more facial muscles. Since any expression that humans can do can be characterized by a concrete combination of Action Units, its automatic recognition is one of the most interest-

1550-5499/15 $31.00 © 2015 IEEE
DOI 10.1109/ICCV.2015.422

ing problems in facial behavior analysis.

The recognition of universal expressions and Action Units can be considered closely related problems. Many psychological studies have empirically shown their strong relation [14]. For instance, Ekman developed the EMFACS dictionary [8], a set of rules mapping Action Unit activation patterns to emotions. Other studies have shown that the expression of a given emotion does not always follow a fixed pattern but that there exist a statistical correlation with concrete Action Unit activations [11, 23].

1.1. Motivation

Action Unit recognition is a challenging problem due to different factors such as illumination changes, pose variations or individual subject differences. One way to advance the field would be by adding larger, and more varied data sets. However, Action Unit annotation is an expensive and laborious task: labeling AUs in one minute of video can require one hour for a specially trained coder. As a consequence, current Action Unit datasets are typically obtained in controlled laboratory conditions and have limitations in terms of positive samples or subject variability. The use of this limited training data for learning AU classifiers can decrease their performance and generalization ability in new testing data. For instance, [30] showed that more reliable smile (Action Unit 12) detectors can be trained using larger datasets collected in naturalistic conditions.

In this work, we ask the following question: *Can we use additional samples labelled with prototypical facial expressions in order to learn better Action Unit classifiers?* Collecting universal facial expression databases is much easier. For instance, the FER2013 Challenge Dataset [10] provides thousands of facial expression samples automatically collected from the Google image search engine. Moreover, facial expression annotations does not require expert coders as in the case of Action Units. Therefore, ground-truth labels for large facial expression datasets are much more easy to obtain compared to Action Units annotations.

1.2. Contributions

Given the previous described motivation, the contributions of the presented work are the following:

- We propose a novel learning framework called Hidden-Task Learning (HTL) that allows to learn a set of Hidden-Tasks when no annotated data is available. For this purpose, HTL exploits prior knowledge about the relation between these Hidden-Tasks and a set of Visible-Tasks for which training data is provided. Additionally, we extend HTL to Semi-Hidden-Task-Learning (SHTL) which is able to use additional training samples belonging to the Hidden-Tasks.
- We show how HTL and SHTL can be used to improve the generalization ability of Action Unit classifiers (Hidden-Tasks) by using additional training data labelled according to prototypical facial expressions (Visible-Tasks). The prior knowledge defining the relation between the AU and Facial Expression recognition tasks is based on empirical results of psychological studies [11]. Even though previous work has used this knowledge for facial expression analysis [25], to the best of our knowledge, this is the first work which exploits it in order to investigate how additional training data of facial expressions can be used to learn better AU classifiers. An overview of our method is provided in Fig.1.
- Performing exhaustive experiments over four different Action Unit databases, our results demonstrate that using SHTL, we can improve AU recognition performance by using additional data from Facial Expression Datasets. In cross-database experiments, HTL generally achieves better performance than standard Single-Task-Learning even when no Action Unit annotations are used. Moreover, SHTL achieves competitive results compared with Transductive Learning approaches which use test data during training in order to learn personalized models for each subject. Our results suggest that the limitation of training data in AU recognition is an important factor which can be effectively addressed with the proposed HTL and SHTL frameworks.

2. Related Work

Action Unit recognition: Most works on AU recognition have focused on proposing different types of facial-descriptors and classification models. Popular descriptors are based on LBP [12], SIFT [3], Active Appearance Models [15] or face-geometry [20] features. On the other hand, different classifiers based on SVM [18], AdaBoost [32] or HMM [26] have been used to recognize Action Units in images or sequences. A review of facial-descriptors and classifiers is out of the scope of this work and related surveys can be found in [34, 4]. However, these approaches do not explicitly face the problem of limited training data in Action Unit recognition. In this work, we show that using simple linear classifiers and standard facial-features, the proposed HTL and SHTL frameworks can increase the generalization ability of AU classifiers (Hidden-Tasks) by just providing additional training samples labelled with facial expressions (Visible-Tasks).

Transductive learning for AU recognition: Individual subject differences suppose one of the main challenges in Action Unit recognition . Recently, [9] have shown that the variability of subjects in the training set plays an impor-

tant role determining the generalization ability of learned models. Therefore, the limited number of subjects in current databases complicates the learning process. In order to address this problem, some works have used Transductive Learning to train personalized AU classifiers by using unlabelled data from the test subject. Chu et al. [3] proposed a method called Selective Transfer Machine. STM learns a penalized SVM by weighting training samples according to their similarity to unlabelled test data. Similarly, Transductive Parameter Transfer [21, 33] learns a mapping from the sample distribution of the test subject to the parameters of a personalized AU classifier. Note that Transductive Learning can be considered an opposite solution to ours. Instead of training specific models for each subject, our approach can use samples from additional subjects present in the facial expressions data in order to learn more generic AU classifiers. Although Transductive Learning approaches have achieved promising results, they are limited in real applications where training classifiers for each subject in testing time is not practical.

Combining AU with Facial Expressions: Exploiting the relation between Action Units and Facial Expressions has been previously explored in the field. Some works have considered to classify expressions by using Action Unit information. For instance, [25] proposed to use a set of rules based on the EMFACS dictionary in order to recognize facial expressions from estimated AU outputs. Similarly, [27] used the Longest Common Subsequence algorithm in order to classify expressions by measuring the similarity between Action Unit patterns in testing and training images. Our work differs from these approaches because we do not use this relation for facial expression recognition but we use it to learn better AU classifiers. Following this idea, some other works have used probabilistic graphical models such as Restricted Boltzmann Machines [29] or Partially-Observed HCRF [2] in order to include facial expression annotations during AU classifiers learning. However, these approaches use samples labelled with both facial expressions and Action Units requiring even more annotation effort. Therefore, they can not be used in order to evaluate how additional training data from facial expression databases can improve Action Unit recognition.

3. HTL and SHTL

Hidden-Task and Semi-Hidden-Task Learning are general purpose frameworks. They can be used in problems where we want to learn a set of Hidden-Tasks for which training data is limited but training samples are easier to obtain from a set of related Visible-Tasks. Note that we consider the set of Hidden and Visible-Tasks disjoint. The use of additional training data from the Visible-Tasks is expected to increase Hidden-Tasks performance. In this section, we formalize the proposed frameworks.

3.1. Hidden-Task Learning

In HTL, we are provided with a training set $\mathbf{X}^v = \{(\mathbf{x}_1^v, \mathbf{y}_1^v), (\mathbf{x}_n^v, \mathbf{y}_n^v), ..., (\mathbf{x}_N^v, \mathbf{y}_N^v)\}$. Each $\mathbf{x}_n \in \mathbb{R}^d$ represents the sample features and $\mathbf{y}_n^v = [y_{n1}^v, y_{nk}^v, ..., y_{nK}^v] \in \{0,1\}^K$ is a vector indicating its label for a set of K binary Visible-Tasks. Using $\mathbf{X}^v$, our goal is to learn a set of T Hidden-Tasks for which training data is not provided.

We denote a Hidden-Task t as a function $\mathbf{h}(\mathbf{x}, \theta_t)$ mapping a feature vector $\mathbf{x}$ to an output according to some parameters θ_t. Given the set of task parameters $\Theta = \{\theta_1, \theta_t, ..., \theta_T\}$, we define the Input-Hidden-Task function:

$$\mathbf{H}(\mathbf{x}, \Theta) = < \mathbf{h}(\mathbf{x}, \theta_1), \mathbf{h}(\mathbf{x}, \theta_t), ..., \mathbf{h}(\mathbf{x}, \theta_T) >^T, \quad (1)$$

mapping $\mathbf{x}$ to a vector containing the outputs of all the T Hidden-Task.

Similarly to Θ, we denote $\Phi = \{\phi_1, \phi_k, ..., \phi_K\}$ as a set of parameters for the K Visible-Tasks. For a given ϕ_k, the Hidden-Visible-Task function $\mathbf{v}(\mathbf{H}(\mathbf{x}_n, \Theta), \phi_k)$ maps $\mathbf{H}(\mathbf{x}_n, \Theta)$ to the output for the Visible-Task k. We assume that Φ can be obtained before the training stage by exploiting prior knowledge about the relation between Hidden and Visible-Task outputs (see Sec. 4.2 for the case of Action Unit and Facial Expressions recognition tasks)

Given the previous definitions, HTL aims to learn the optimal Hidden-Task parameters Θ by minimizing:

$$\min_{\Theta, \mathbf{X}^v} \; \mathcal{L}^v(\Theta, \mathbf{X}^v) + \beta \mathcal{R}(\Theta). \quad (2)$$

Here, $\mathcal{R}(\Theta)$ refers to a regularizer over the parameters Θ preventing over-fitting, $\mathcal{L}^v$ is the empirical-risk over the Visible-Task training set $\mathbf{X}^v$ defined in Eq. 3 and ℓ can be defined as any classification loss function. The parameter β controls the impact of the regularization term.

$$\mathcal{L}^v(\Theta, \mathbf{X}^v) = \frac{1}{NK} \sum_{n=1}^{N} \sum_{k=1}^{K} \ell(\mathbf{v}(\mathbf{H}(\mathbf{x}_n^v, \Theta), \phi_k), y_{nk}^v), \quad (3)$$

Note that HTL shares some relation with weakly-supervised structured learning [28]. In our case, the goal is to learn a set of Hidden-Tasks predicting a latent structured output $\mathbf{H}(\mathbf{x}, \Theta)$ by using only the visible weak-labels y^v. As discussed, HTL is able to solve this problem by pre-defining the relation between Hidden and Visible-Tasks based on prior knowledge.

3.2. Semi-Hidden Task Learning

In SHTL, we assume that additional training data for the Hidden-Tasks is provided. Similarly to $\mathbf{X}^v$, we denote $\mathbf{X}^h = \{(\mathbf{x}_1^h, \mathbf{y}_1^h), (\mathbf{x}_m^h, \mathbf{y}_m^h), ..., (\mathbf{x}_M^h, \mathbf{y}_M^h)\}$ as a training set of M samples where $\mathbf{y}_m^h \in \{0,1\}^T$ indicates the sample class label for each Hidden-Task t. Following the definitions in the previous section, now we are interested in

learning the optimal parameters Θ by minimizing:

$$\min_{\Theta} \ (1-\alpha)\mathcal{L}^h(\Theta, \mathbf{X}^h) + \alpha\mathcal{L}^v(\Theta, \mathbf{X}^v) + \beta\mathcal{R}(\Theta) \quad (4)$$

where $\mathcal{L}^h(\mathbf{\Theta}, \mathbf{X}^h)$ represents the empirical-risk function over the Hidden-Task training set $\mathbf{X}^h$:

$$\mathcal{L}^h(\mathbf{\Theta}, \mathbf{X}^h) = \frac{1}{MT}\sum_{m=1}^{M}\sum_{t=1}^{T}\ell(\mathbf{h}(\mathbf{x}_m^h, \theta_t), y_{mt}^h). \quad (5)$$

The parameter $\alpha \in [0, 1]$ controls the trade-off between the minimization of the Hidden-Task and Visible-Task losses. Concretely, note that when $\alpha = 1$ the minimization is the same as HTL. In contrast, when $\alpha = 0$, SHTL is equivalent to learning the Hidden-Tasks without taking into account the Visible-Tasks training data, i.e., traditional Single-Task Learning (STL). Therefore, SHTL can be considered a generalization of both HTL and STL.

An interesting interpretation of SHTL is to understand the term $\alpha\mathcal{L}^v(\Theta, \mathbf{X}^v)$ in Eq. 4 as a regularization function. Concretely, it penalizes cases where the Hidden-Task-outputs in x^v are not coherent with its label y^v according to the known relation between Hidden and Visible tasks. To the best of our knowledge, this is a novel idea which can be useful in different problems than AU recognition where training data is limited but samples are easier to annotate for a set of related tasks.

4. From universal emotions to Action Units

The use of HTL and SHTL allow us to evaluate how larger training sets can improve Action Unit recognition. Using the relation between AUs and universal facial expressions, we can learn Action Unit classifiers (Hidden-Tasks) by training them using additional samples labelled with prototypical facial expressions (Visible-Tasks). As previously discussed, the use of additional training data is expected to improve classifier performance by increasing their generalization ability. Following, we describe how we apply both HTL and SHTL frameworks to this particular problem.

4.1. Defining HTL and SHTL for AU recognition

For HTL, we assume that we are only provided with a facial expressions training set $\mathbf{X}^v$ composed by N samples. Each $\mathbf{x}_n^v \in \mathbb{R}^D$ is a facial-descriptor extracted from a face image and $\mathbf{y}_n^v \in \{0, 1\}^K$ indicates its expression label. In this case, K=7 because we consider the 6 universal facial expressions plus the neutral face. In SHTL, we are also provided with an Action Unit training set $\mathbf{X}^h$ of M samples. The label vector $\mathbf{y}_m^h \in \{0, 1\}^T$ indicates what Action Units are present in $\mathbf{x}_m^h$. Note that T refers to the number of Action Units considered.

The Hidden-Task parameters Θ are defined as $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_t, ..., \mathbf{a}_T]$. Each $\mathbf{a}_t \in \mathbb{R}^D$ is a linear classifier and the Hidden-Task function $\mathbf{h}(\mathbf{x}, \mathbf{a_t})$:

$$p_t(\mathbf{x}) = \mathbf{h}(\mathbf{x}, \mathbf{a_t}) = (1 + \exp(-\mathbf{a}_t^T\mathbf{x}))^{-1}, \quad (6)$$

represents the probability of the Action Unit t given an input feature $\mathbf{x}$ modelled with a sigmoid function.

Now we define $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_k, ..., \mathbf{e}_K]$ as the set of Visible-Task parameters Φ. Each $\mathbf{e}_k \in R^T$ is also a linear classifier mapping the set of T Action Unit probabilities to an output for the facial expression k. Concretely, the Hidden-Visible-Task function $\mathbf{v}(\mathbf{H}(\mathbf{x}, \mathbf{A}), \mathbf{e}_k)$ is defined as:

$$p_k(\mathbf{x}) = \mathbf{v}(\mathbf{H}(\mathbf{x}, \mathbf{A}), \mathbf{e}_k) = \frac{\exp(\mathbf{e}_k^T\mathbf{H}(\mathbf{x}, \mathbf{A}))}{\sum_{r=1}^{K}\exp(\mathbf{e}_r^T\mathbf{H}(\mathbf{x}, \mathbf{A}))} \quad (7)$$

and denotes the probability of the facial expression k given the set of Action Unit outputs $\mathbf{H}(\mathbf{x}, \mathbf{A})$.

Given the previous definitions, the Visible-Task Loss is defined as the cross-entropy error function over the Facial-Expression-Recognition tasks as:

$$\mathcal{L}^v(\mathbf{A}, \mathbf{X}^v) = \frac{-1}{NK}\sum_{n=1}^{N}\sum_{k=1}^{K} y_{nk}^v \ln(p_k(\mathbf{x}_n^v)). \quad (8)$$

Similarly, the Hidden-Task Loss is defined as the log-loss function over the set of Action Unit classification tasks:

$$\mathcal{L}^h(\mathbf{A}, \mathbf{X}^h) = \frac{-1}{MT}\sum_{m=1}^{M}\sum_{t=1}^{T} y_{mt}^h \ln(p_t(\mathbf{x}_m^h)) + (1 - y_{mt}^h)(1 - \ln(p_t(\mathbf{x}_m^h))) \quad (9)$$

Finally, we use standard L2-regularization $\frac{1}{2}\sum_{t=1}^{T}||\mathbf{a}_t||_2^2$ for the Hidden-Task parameters regularizer $\mathcal{R}(\mathbf{A})$.

4.2. Training the AU-Emotions Tasks Function

One of the key points in HTL and SHTL is how to obtain the Visible Tasks parameters Φ before training. In our case, we need to obtain a set of linear classifiers $\mathbf{E} = [\mathbf{e_1}, \mathbf{e_2}, ..., \mathbf{e_K}]$ mapping Action Unit activations to an output for each facial expression. For this purpose, we exploit the empirical results reported in [11, 23]. In these psychological studies, a set of actors were recorded while they interpreted situations involving the six universal basic emotions defined by Ekman. Then, AU annotations were obtained for each video according to the Facial Action Coding System and Action Unit frequencies for each emotion were computed (see Fig. 2(a)). More details can be found in the original publications.

We use these empirical results in order to train the Visible-Task classifiers $\mathbf{E}$ as follows. For each emotion, we generate a large number of random samples $\mathbb{R} \in [0, 1]^T$ assuming that the probability of an AU activation follows a Bernoulli distribution according to its mean frequency in

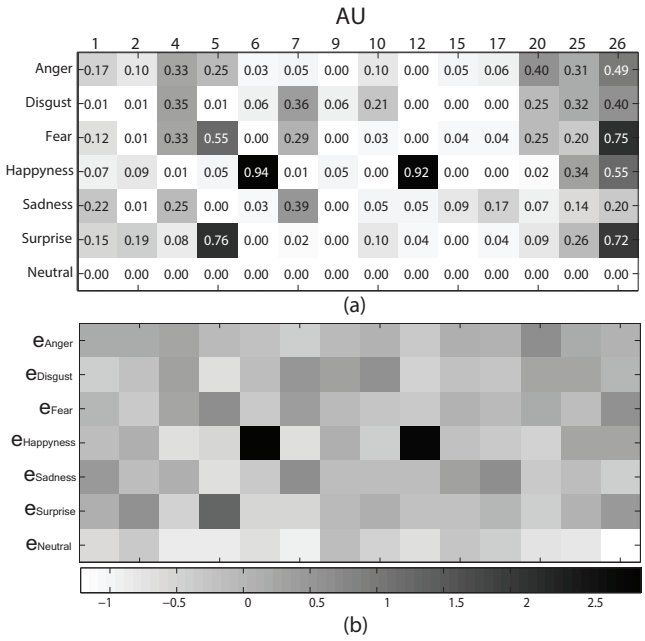

Figure 2: (a) Action Unit activation probability for each emotion obtained in [11]. In Action Unit 20, we have used the results obtained in [23] for Anger and Fear emotions [1]. (b) Trained linear classifiers $\mathbf{E}$ mapping AU activations to emotions. See text for details.

Fig. 2. For each sample dimension, we assign a random value between 0 and 0.5 if the AU is activated and between 0.5 and 1 otherwise. Intuitively, these samples are vectors simulating possible Action Unit activations for each type of emotion according to Eq. 6. Finally, we train a linear multiclass-SVM using the generated samples in order to obtain the classifiers $[\mathbf{e_1}, \mathbf{e_2}, ..., \mathbf{e_K}]$. Obtained coefficients for each $\mathbf{e_k}$ are shown in Fig. 2(b).

4.3. Optimization

According to Eq. 4, we need to solve:

$$\min_{\mathbf{A}} \ (1-\alpha)\mathcal{L}^h(\mathbf{A}, \mathbf{X}^h) + \alpha\mathcal{L}^v(\mathbf{A}, \mathbf{X}^v) + \beta\mathcal{R}(\mathbf{A}) \quad (10)$$

in order to obtain the set of optimal Action Unit classifiers $\mathbf{A}$. For this purpose, we follow a gradient-descent approach. Concretely, we use the L-BFGS Quasi-Newton method [1] which provides a higher-convergence rate than first order gradient-descent approaches and approximates the Hessian matrix with a low-rank compact form. The gradient of $\mathcal{R}(\mathbf{A})$, $\mathcal{L}^v(\mathbf{A}, \mathbf{X}^v)$ and $\mathcal{L}^h(\mathbf{A}, \mathbf{X}^h)$ w.r.t each vector $\mathbf{a}_t$ are:

$$\nabla\mathcal{L}^v = \frac{-1}{NK}\sum_{n=1}^{N}\sum_{k=1}^{K} y^v_{nk}(e^{(t)}_k - \sum_{s=1}^{K} p_{nk}e^{(t)}_s)p_{nt}(1-p_{nt})\mathbf{x}^v_n$$

$$\nabla\mathcal{R}(\mathbf{A}) = \mathbf{a}_t \ , \ \nabla\mathcal{L}^h = \frac{-1}{MT}\sum_{n=1}^{M}(y^v_{mt} - p_{mt})\mathbf{x}^h_m. \quad (11)$$

For shorter notation, we use $p_{nk} = p_k(\mathbf{x}^v_n)$ and $p_{mt} = p_t(\mathbf{x}^h_m)$. $e^{(t)}_k$ is a scalar corresponding to the dimension t of the vector $\mathbf{e}_k$

5. Experiments

In Sec. 5.1 and Sec. 5.2 we describe the different datasets and facial features used in our experiments. In the following sections, we discuss the different experiments and obtained results evaluating the proposed HTL and SHTL frameworks for Action Unit recognition.

5.1. Databases

Action Unit Databases: We have used four different Action Unit databases widely used in the literature: the Extended Cohn-Kanade (CK+) [16], the GEMEP-FERA [24], the UNBC-McMaster Shoulder Pain Expression [17] and the DISFA [19] datasets. CK+ contains 593 sequences of different subjects performing posed Action Units from the neutral face to the AU appex. Same as [3], we use the first frame as a negative sample and the last third frames as positive ones. The GEMEP-FERA data set contains 87 recordings of 7 different actors simulating a situation eliciting a concrete emotion. The UNBC database contains a set of 200 videos of 25 different patients undergoing shoulder pain. These patients were recorded while doing different types of arm movements. Finally, the DISFA dataset contains 27 videos of different subjects watching Youtube videos choosen in order to elicit different types of emotions. AU annotations are provided for each frame. Note that these four data-sets include posed, acted and spontaneous facial behavior. In our experiments, we have considered the recognition of Action Units 1,2,4,5,6,7,9,10,12,15,17,20,25 and 26 which include the 7 most frequent lower and upper AUs over the four datasets.

Facial expression data: In order to obtain a large number of variated facial expression images, we have collected samples from different datasets annotated with the 6 universal emotions (anger, disgust, happiness, sadness, fear and surprise) plus the neutral face. From the Bosphorous Database [22], we have used a set of 752 frontal face images from 105 different subjects. From the Radboud Faces [13] Database, we have obtained 469 frontal face images from 67 subjects. Finally, with a similar process as followed in the FER2013 Challenge [10], we have automatically collected thousands of images from Google and Bing search engines [2]. For this purpose, we used a set of 70 composed

[1] As reported in [23], we observed that AU20 is also present in some Anger and Fear expression images. However, it is not reflected by the empirical results obtained in [11]

[2] We have considered to collect our own database because the provided images in [10] have a low resolution (48x48) and the annotations are very noisy. It will be made available upon request for the research community

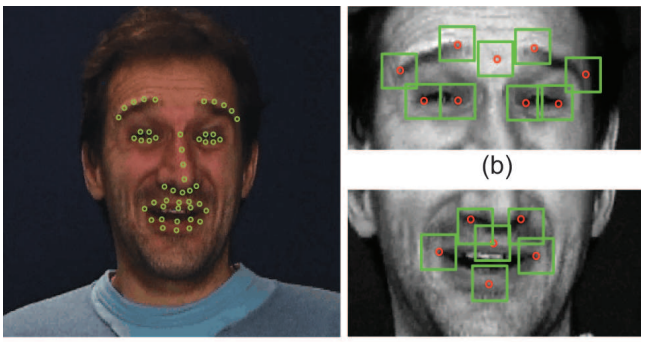

Figure 3: Facial-descriptors extracted for the upper and lower part of the face. (a) Original image with the set of 49 landmarks points obtained with [31]. (c,d) Aligned face image and local patches used to extract the SIFT features composing the lower and upper facial descriptors.

		AUC				F1			
Test	Train	SVM	STL	SHTL	HTL	SVM	STL	SHTL	HTL
CK+	UNBC	75.7	78.2	**81.7**	78.3	40.2	43.4	**49.2**	47.2
	FERA	76.6	75.5	**83.4**	80.6	41.6	38.2	**54.7**	51.6
	DISFA	83.4	84.3	**86.1**	83.7	52.8	54.8	**60.1**	56.7
UNBC	CK+	68.2	68.4	**69.7**	**69.7**	**16.9**	**16.9**	15.8	15.6
	FERA	63.8	65.2	**70.0**	69.7	12.9	13.6	**15.7**	15.6
	DISFA	67.1	67.4	**69.2**	68.8	16.3	16.2	**18.0**	16.4
FERA	CK+	70.8	70.8	**72.4**	68.0	43.1	41.3	**44.7**	40.9
	UNBC	67.5	69.4	**71.5**	70.0	42.2	40.5	42.7	**45.5**
	DISFA	70.4	71.3	**72.4**	68.9	44.2	44.3	**45.0**	39.7
DISFA	CK+	71.7	72.6	**76.0**	74.4	30.8	33.5	**39.1**	36.1
	UNBC	69.7	70.3	74.0	**76.7**	32.4	35.7	43.5	**45.4**
	FERA	68.6	70.3	**75.6**	74.4	25.2	25.5	**38.5**	36.1
	Avg.	71.1	72.0	**75.2**	73.6	33.2	33.7	**38.9**	37.3

Table 1: Average AU recognition performance obtained with SVM, STL, SHTL and HTL in the set of twelve cross-database experiments. Colors illustrate the different approaches ordered according to their performance.

queries such as "*sad man*","*disgusted woman*" or "*happy face*". Then, images which did not correspond to their emotion query were filtered by a non-expert annotator. Overall, we have collected 3437 facial expression images with a large variety of subjects, illuminations and other factors. In order to test labels reliability, an additional coder repeats the same process in 300 images for each facial expression (2100 images in total). The observed inter-coder agreement was 0.89 with a Cohen's Kappa coefficient of 0.78. Finally, we have augmented the number of samples by flipping each image around the vertical axis.

5.2. Facial features

As we have explained in Sec. 4.1, we consider a sample $\mathbf{x}$ as a facial-descriptor obtained from a given face image. Before extracting it, we follow a face-alignment process. Firstly, we automatically detect 49 facial-landmarks with the method described in [31]. Secondly, we compute an affine transformation aligning the obtained points with a mean shape. Finally, we apply the transformation to the image and crop the face region (see Fig. 3(a)-(b)). From the obtained aligned face, we extract two facial-descriptors from the upper and lower half parts of the face similar to [3]. The use of two different features from both parts is motivated by the fact that different Action Units are localized in concrete face areas such as eyes, eyebrows, mouth, etc... Therefore, it is convenient that AU classifiers use one of these descriptors depending on the localization of its corresponding AU. Concretely, we extract a set of SIFT descriptors from local patches centered in a subset of the landmarks (see Fig. 3(c)-(d)). Features for each part are concatenated in order to form the final lower and upper facial-descriptors.

5.3. Cross-Databases experiments

We evaluate how HTL and SHTL can be used to improve the generalization ability of AU classifiers by providing additional facial expression samples during training. For this purpose, we have designed a set of cross-database experiments where one Action Unit dataset is used for training and one for testing. In contrast to most works which train and test on the same data-set, a cross-database validation provides more information about how AU classifiers generalize to new subjects and other factors.

Under this setting, we compare the performance of HTL and SHTL with standard Single-Task-Learning (STL). Remember that we refer to STL when only Action Unit training data is used. On the other hand, HTL uses only samples from the Facial Expression dataset and SHTL uses both. As explained in Sec. 3.2, these three approaches are generalized by the proposed SHTL framework by changing the α value in Eq. 10. We use α=0 for STL, α=1 for HTL and α=0.5 for SHTL. As a baseline, we also evaluate the performance of a linear SVM classifier trained independently for each AU. Note that SVM can also be considered a Single-Task-Learning approach with a different loss function than our STL. The regularization parameter β has been obtained by cross-validation over the training set. Table 1 shows the obtained average AUC and F1-score for the considered set of 14 Action Units[3]. Detailed results for each independent AU are provided in supplementary material. [4]

HTL vs STL and SVM: Comparing HTL to STL and SVM, we can observe that HTL achieve comparable or bet-

[3]Only AUs available in the training dataset are used to compute results. HTL and SHTL can learn AU classifiers even when no AU samples are provided in the training set. However, for a fair comparison with STL and SVM, we do not consider these cases to evaluate performance. This explains HTL performance differences on the same test set.

[4]http://cmtech.upf.edu/research/projects/shtl

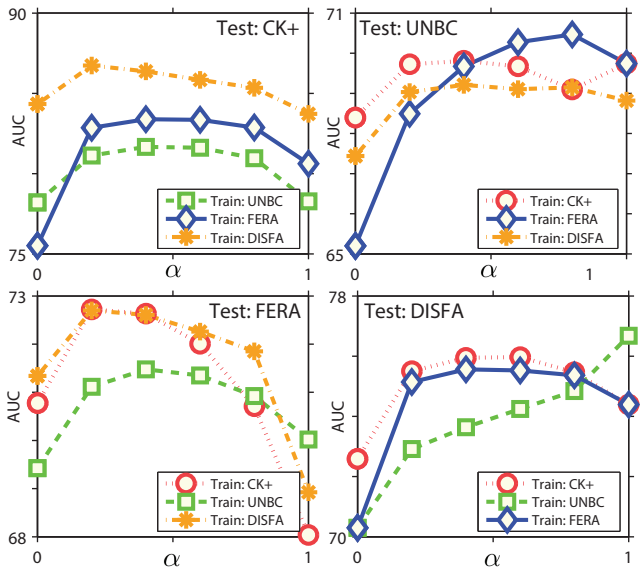

Figure 4: Average AU recognition performance in the cross-database experiments varying the α parameter in the range between 0 and 1. See text for details.

ter performance in terms of average AUC and F1 for most of the cross-database experiments. It could seem surprising because HTL does not use any Action Unit annotation during training. However, it confirms our hypothesis that the limited training data of current AU datasets can decrease the quality of learned models. In contrast, HTL uses richer facial expression data which increases its generalization ability over different datasets. Additionally, notice that STL and SVM achieves similar average performance. This can be explained because both are Single-Task-Learning approaches which only use the Action Unit data for training.

SHTL vs STL and HTL: Comparing SHTL with the other approaches, we can observe that SHTL achieves superior performance in most cases. These can be explained because SHTL is able to combine information from the AU and Facial Expression training samples. Analyzing the performance for each AU independently, the results show some variations depending on each experiment. However, SHTL generally outperforms either HTL or STL. Again, it shows the advantages of using SHTL in order to combine both AU and facial expression training data information.

Evaluating the effect of α parameter: Previously, we have fixed the α parameter of SHTL to 0.5. This provides a balanced trade-off between Hidden (Action Units) and Visible-Task (Facial Expressions) losses. However, different values for α are also possible. In order to evaluate the impact of the α parameter, we have run the same set of experiments fixing it to different values in the range between 0 to 1. As Figure 4 shows, optimal performance is generally obtained with α between 0 and 1 which combines information from AU and Facial Expression databases (SHTL).

	AUC				F1			
Train	SVM	STL	SHTL	HTL	SVM	STL	SHTL	HTL
CK+	90.6	91.2	**91.7**	80.6	68.5	68.6	**68.9**	51.7
UNBC	75.3	78.2	**78.8**	69.7	22.7	21.3	**27.1**	15.6
FERA	66.7	66.9	**73.4**	68.0	46.8	48.7	**51.9**	40.9
DISFA	79.6	81.2	**81.5**	74.4	37.6	40.5	**42.9**	36.1
Avg.	78.0	79.4	**81.3**	73.2	43.9	44.8	**47.7**	36.1

Table 2: Average Action Unit recognition performance obtained with SVM,STL,SHTL and HTL in single-dataset experiments. Colors illustrate the different approaches ordered according to their performance in each experiment.

We have shown that by using HTL and SHTL, the use of additional training data labelled with prototypical facial expressions improves the generalization ability of learned AU classifiers. Note that we are using simple linear classifiers and standard facial-features. However, these frameworks are flexible enough to be used with any kind of facial-descriptors or base classifiers.

5.4. Single-database experiments

Although cross-database experiments are useful to evaluate the generalization ability of learned models, it is reasonable to ask how SHTL and HTL performs in Action Unit data which have been obtained in similar conditions. In this experiment, we evaluate the previously used methods with a leave-one-subject strategy over the same dataset. Note that this setting is similar to the commonly used in the literature. In this case, for SHTL we have set $\alpha = 0.25$ in order to give more importance to the Hidden-Task loss (Action Unit data). Moreover, for SVM, STL and SHTL we have optimized the classification threshold using the Action Unit training samples during cross-validation. [5]

Figure 2 shows the obtained results. Under this setting, HTL achieves the worst performance. However, it was expected since the problem of generalizing to data taken in different conditions is mitigated in this case. SHTL achieves slightly better AUC than STL and SVM in all the cases and a more significant improvement in terms of the F1-score. Therefore, even when data is taken in similar conditions, the use of additional facial expression samples is beneficial. One of the main factors that could explain SHTL improvement is that current Action Unit databases are limited in terms of subject variability. Therefore, SHTL can learn more generic AU classifiers by using training samples from additional subjects present in the facial expressions data. One point that supports that conclusion is that SHTL obtains a significant improvement over the FERA dataset which is the most limited in terms of subjects. In contrast,

[5] Worst results were observed optimizing the threshold in cross-database experiments.

this improvement is less significative in the CK+ dataset which has the larger number of subjects.

5.5. Comparison with related work: Tranductive Learning

In this experiment, we compare SHTL with state-of-the-art transductive learning approaches for AU recognition: STM [3], TPT [21] and SVTPT [33]. As we have discussed in Sec. 2, these methods use unlabelled data during training in order to learn personalized models for each test subject. In contrast, SHTL is trained with additional facial expressions data which increases its generalization ability to new subjects. We have used similar features and followed the same experimental-setup in order to compare our results with the reported in the cited works. We have retrained the classifiers $\mathbf{e}_k$ (Sec. 4.2) using only the subset of 8 AUs evaluated in STM. They also include the 6 AUs used in TPT and SVTPT works. Again, the α parameter of SHTL has been set to 0.25.

Table 3 shows the obtained results. As the reader can observe, SHTL achieves competitive performance compared with transductive learning approaches. Concretely, SHTL obtains better AUC in all cases and similar F1-score over the CK+ dataset. Only STM significantly outperforms the F1-score of SHTL in the FERA dataset. However, it is worth mentioning that Transductive Learning models need to be trained for each subject during testing and requires sufficient samples to correctly estimate the test distribution. In contrast, SHTL just needs to learn a single generic classifier by using the additional facial expression data. Therefore, SHTL is more useful in real applications where training Action Unit classifiers for each subject during testing is not feasible (e.g. online detection of Action Units in video streams).

		SHTL	STM [3]	TPT [21]	SVTPT [33]
AUC	FERA	**76.2**	74.5	-	-
	CK+ (8 AUs)	**93.4**	91.3	-	-
	CK+ (6 AUs)	**93.9**	90.1	91.3	92.7
F1	FERA	55.9	**59.9**	-	-
	CK+ (8 AUs)	**76.5**	**76.6**	-	-
	CK+ (6 AUs)	78.8	74.8	76.8	**79.1**

Table 3: SHTL performance and results reported by state-of-the-art transductive Learning approaches for Action Unit recognition on CK+ and FERA datasets.

6. Conclusions

In this paper, we have investigated how additional training data annotated with universal facial expressions can improve the generalization ability of Action Unit classifiers. For this purpose, we have proposed the Hidden and Semi-Hidden Task Learning frameworks able to learn a set of Hidden-Tasks (Action Units) when training data is limited or even not available. These frameworks are able to exploit prior knowledge about the relation between these Hidden-Tasks and a set of Visible-Tasks (Facial Expressions).

Exhaustive experiments have shown that HTL and SHTL improve the generalization ability of Action Unit classifiers by using training data from a large facial expression database. Surprisingly, HTL generally achieves better performance than standard Single-Task Learning in cross-database experiments without using any Action Unit annotation. Moreover, we have also shown the advantages of combining AU and Facial Expressions data information with SHTL. Despite that most existing work on AU recognition has focused on proposing facial features or classification methods, our results suggest that the limitation of training data in AU recognition is an important factor which has been largely overlooked. The proposed HTL and SHTL frameworks can address this problem by using additional training data annotated with facial expression labels which are much easier to obtain.

As a future work, we plan to study how to adapt the Visible Task Layer during training by using only the pre-trained parameters as a prior. It could allow SHTL to correct possible inaccuracies of the empirical studies relating Facial Expressions with Action Unit occurrences. Finally, we consider that HTL and SHTL are general purpose frameworks which could be also useful in other problems where the lack of annotated training data is a challenge.

Acknowledgements

This paper is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 645012. Adria Ruiz and Xavier Binefa would also like to acknowledge Spanish Government to provide support under grants CICYT TIN2012-39203 and FPU13/01740. Joost van de Weijer acknowledges Project TIN2013-41751 of the Spanish Ministry of Science and the Generalitat de Catalunya Project under Grant 2014-SGR-221.

References

[1] R. H. Byrd, J. Nocedal, and R. B. Schnabel. Representations of quasi-newton matrices and their use in limited memory methods. *Mathematical Programming*, 1994. 5

[2] K.-Y. Chang, T.-L. Liu, and S.-H. Lai. Learning partially-observed hidden conditional random fields for facial expression recognition. In *Proc. Computer Vision and Pattern Recognition*, 2009. 3

[3] W.-S. Chu, F. De la Torre, and J. F. Cohn. Selective transfer machine for personalized facial action unit detection. *Proc. Computer Vision and Pattern Recognition*, 2013. 2, 3, 5, 6, 8

[4] F. De la Torre and J. F. Cohn. Facial expression analysis. In *Visual Analysis of Humans*. 2011. 2

[5] S. Du, Y. Tao, and A. M. Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 2014. 1

[6] P. Ekman and W. Friesen. Facial Action Coding System: A Technique for the Measurement of Facial Movement. *Consulting Psychologists Press*, 1978. 1

[7] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 1971. 1

[8] W. V. Friesen and P. Ekman. Emfacs-7: Emotional facial action coding system. *Unpublished manuscript, University of California at San Francisco*, 1983. 2

[9] J. Girard, J. Cohn, and F. De La Torre. How much training data for facial action unit detection? *International Conference on Automatic Face and Gesture Recognition*, 2015. 2

[10] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *Advances in Neural Information Processing Systems*, 2013. 2, 5

[11] P. Gosselin, G. Kirouac, and F. Y. Doré. Components and recognition of facial expression in the communication of emotion by actors. *Journal of personality and social psychology*, 1995. 2, 4, 5

[12] B. Jiang, M. F. Valstar, and M. Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. In *International Conference on Automatic Face and Gesture Recognition*, 2011. 2

[13] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. van Knippenberg. Presentation and validation of the radboud faces database. *Cognition and Emotion*, 2010. 5

[14] M. Lewis, J. M. Haviland-Jones, and L. F. Barrett. Handbook of emotions. chapter 13. 2010. 2

[15] P. Lucey, J. Cohn, S. Lucey, S. Sridharan, and K. M. Prkachin. Automatically detecting action units from faces of pain: Comparing shape and appearance features. In *Proc. Computer Vision and Pattern Recognition Workshops*, 2009. 2

[16] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Proc. Computer Vision and Pattern Recognition Workshops*, 2010. 5

[17] 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 *International Conference on Automatic Face and Gesture Recognition*, 2011. 5

[18] M. H. Mahoor, S. Cadavid, D. S. Messinger, and J. F. Cohn. A framework for automated measurement of the intensity of non-posed facial action units. In *Proc. Computer Vision and Pattern Recognition Workshops*, 2009. 2

[19] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. Disfa: A spontaneous facial action intensity database. *Transactions on Affective Computing*, 2013. 5

[20] M. Pantic and I. Patras. Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. *Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2006. 2

[21] 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 *ACM Multimedia*, 2014. 3, 8

[22] A. Savran, N. Alyüz, H. Dibeklio\uglu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun. Bosphorus database for 3D face analysis. In *Biometrics and Identity Management*, pages 47–56. Springer, 2008. 5

[23] K. R. Scherer and H. Ellgring. Are facial expressions of emotion produced by categorical affect programs or dynamically driven by appraisal? *Emotion*, 2007. 2, 4, 5

[24] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer. Meta-analysis of the first facial expression recognition challenge. *Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2012. 5

[25] M. F. Valstar and M. Pantic. Biologically vs. logic inspired encoding of facial actions and emotions in video. In *Multimedia and Expo, 2006 IEEE International Conference on*, 2006. 2, 3

[26] M. F. Valstar and M. Pantic. Fully automatic recognition of the temporal phases of facial actions. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 42(1):28–43, 2012. 2

[27] S. Velusamy, H. Kannan, B. Anand, A. Sharma, and B. Navathe. A method to infer emotions from facial action units. In *International Conference on Acoustics, Speech and Signal Processing*, 2011. 3

[28] A. Vezhnevets, V. Ferrari, and J. M. Buhmann. Weakly supervised structured output learning for semantic segmentation. In *Proc. Computer Vision and Pattern Recognition*, 2012. 3

[29] Z. Wang, Y. Li, S. Wang, and Q. Ji. Capturing global semantic relationships for facial action unit recognition. In *Proc. IEEE Int. Conf. on Computer Vision*. 3

[30] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan. Toward practical smile detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009. 2

[31] X. Xiong and F. De la Torre. Supervised descent method and its application to face alignment. In *Proc. Computer Vision and Pattern Recognition*, 2013. 6

[32] P. Yang, Q. Liu, and D. N. Metaxas. Boosting coded dynamic features for facial action units and facial expression recognition. In *Proc. Computer Vision and Pattern Recognition*, 2007. 2

[33] G. Zen, E. Sangineto, E. Ricci, and N. Sebe. Unsupervised domain adaptation for personalized facial emotion recognition. *International Conference on Multimodal Interaction*, 2014. 3, 8

[34] Z. Zeng, M. Pantic, G. Roisman, and T. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009. 2