A Framework for Joint Estimation and Guided Annotation of Facial Action Unit Intensity

Robert Walecki*, Ognjen Rudovic*, Maja Pantic*‡, Vladimir Pavlovic†, Jeffrey F. Cohn±
*Department of Computing, Imperial College London, UK
† Department of Computer Science, Rutgers University, USA
‡ EEMCS, University of Twente, The Netherlands
± Department of Psychology, University of Pittsburgh, PA, USA
± The Robotics Institute, Carnegie Mellon University, PA, USA
{r.walecki14,o.rudovic,m.pantic}@imperial.ac.uk, vladimir@cs.rutgers.edu, jeffcohn@pitt.edu

Abstract

Manual annotation of facial action units (AUs) is highly tedious and time-consuming. Various methods for automatic coding of AUs have been proposed, however, their performance is still far below of that attained by expert human coders. Several attempts have been made to leverage these methods to reduce the burden of manual coding of AU activations (presence/absence). Nevertheless, this has not been exploited in the context of AU intensity coding, which is a far more difficult task. To this end, we propose an expert-driven probabilistic approach for joint modeling and estimation of AU intensities. Specifically, we introduce a Conditional Random Field model for joint estimation of the AU intensity that updates its predictions in an iterative fashion by relying on expert knowledge of human coders. We show in our experiments on two publicly available datasets of AU intensity (DISFA and FERA2015) that the AU coding process can significantly be facilitated by the proposed approach, allowing human coders to faster make decisions about target AU intensity.

1. Introduction

Human facial expressions are typically described in terms of variation in configuration and intensity of facial muscle actions defined using the Facial Action Coding System (FACS) [5]. Specifically, the FACS defines a unique set of 33 atomic non-overlapping facial muscle actions named Action Units (AUs) [17]. It also defines rules for scoring the intensity of each AU in the range from absent to maximal intensity on a six-point ordinal scale. Using FACS, nearly any anatomically possible facial expression can be described by decomposing it into specific AUs and their intensities. This is considered the most objective way of describing human facial behaviour, and it has found a variety of applications in behavioural sciences and psychology (e.g., studies on emotion expression), and computer analysis of facial expressions (e.g., pain monitoring [12]).

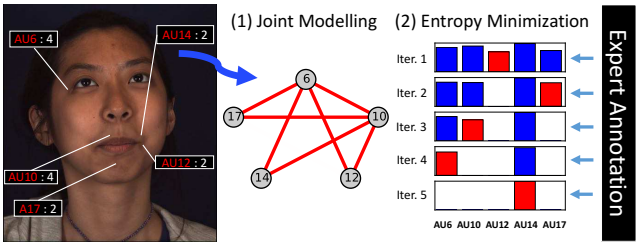


Figure 1: The guided AU intensity annotation. First, the AUs intensities are jointly modeled using a CRF-based classifier. Secondly, the model predictions are used to assist a human coder in an iterative fashion: the coder scores the intensity of an AU (depicted in red) that is the easiest. Then, this labelling is used to obtain more accurate predictions for the remaining AUs (depicted in blue), thus, providing the human coder with more confident suggestions for the intensity of the remaining AUs. This is repeated until all AUs are scored.

The coding of AU intensity can be done manually or automatically, or by combining these two approaches. Manual coding of AUs is typically performed by FACS certified human coders. Apart from the fact that there are no many experienced FACS coders, manual coding of AUs is highly expensive and time-consuming. To illustrate the coding process, a human coder goes through each image frame of a face video, and identifies the intensity level of each AU. To this end, the coder applies the intensity coding rules defined by FACS. In addition to the facial appearance changes being very subtle from one intensity level to another, a different criteria for scoring AU intensity may apply when AUs occur

in combination than when they occur alone. For instance, the criteria for intensity scoring of AU7 (lid tightener) are changed significantly if AU7 appears with a maximal intensity of AU43 (eye closure), since this combination changes the appearance as well as timing of these AUs [5]. Furthermore, co-occurring AUs can be non-additive, in the case of which one AU masks another, or a new and distinct set of appearances is created [5]. Also, some AUs are often activated together, *e.g*. AU12 and AU6 in the case of smiles, but with different intensities depending on the type of smile (e.g., genuine vs. posed). To reduce the efforts, typically, a subset of the most occurring AUs is selected for coding (e.g., 12 AUs in DISFA database [20]). Yet, this process is still tedious and error-prone due to the difficulty of discerning intensities of multiple AUs [18]. Moreover, since a face video is recorded at 25fps at least, manual coding of intensity of AUs can become prohibitively expensive. Finally, to validate the annotations, usually two independent annotators are asked to do the coding, until an acceptable inter-observer reliability is achieved [8]. This makes the whole process even more labour intensive.

To reduce the burden of manual coding of AUs and, in particular, their intensity, various methods have been proposed to automate this process. This has become possible mainly due to the recent advances in computer vision and machine learning, and in particular, the affective computing field. For detailed description of the steps in automated estimation of AUs, the reader is referred to [3]. The existing methods for automated coding of AUs can be divided into those that output either binary (presence vs. absence) [16, 2, 18, 29, 6] or ordinal (intensity) [17, 20, 23, 12, 11] labels. In this work, we focus on the latter as it poses a more challenging problem. The methods for automated estimation of AU intensity can be divided into those that perform static [17, 20, 12, 11] vs. dynamic [23] estimation. While the former focus mainly on engineering of efficient image features for the target task, the latter exploit the temporal unfolding of the AU intensity. Note, however, that most of the existing methods perform the intensity estimation independently for each AU. Only recently, several methods for joint estimation of the AU intensity have been proposed [24, 15, 13]. The main motivation for this approach is that by joint modeling of AUs, the resulting AU intensity classifiers are more robust to the (highly) imbalanced intensity levels (within and between AUs), and non-additive AU combinations. Also, depending on the target context (e.g., emotion expression), different AUs and their intensity levels are more likely to occur together (e.g., AU6&12 in the case of genuine smiles). Joint modeling of AUs is also an attempt to simulate the human (underlying) reasoning during the coding process: even though the humans code each AU separately, they use the contextual information (the whole face) to narrow down the possible AU combinations and their intensity levels that can occur simultaneously in a face image. Thus, knowing the context can reduce the coding time by humans, and also reduce uncertainty in automated coding of AU intensity.

The automatic methods in [24, 15] perform a two stage joint modeling of AU intensity. In [24], the scores of the pre-learned regressors based on Support Vector Regression [1] are fed into Markov Random Field trees [1], used to model dependencies of AU subsets. Similarly, [15] models AU dependencies using a Dynamic Bayesian Network (DBN), applied to the intensity scores of the AU-specific spectral regressors. A more recent approach for joint modeling of the AU intensity formulates a generative MRF model, called Latent Trees (LT) [13]. In contrast to [24, 15], this method can deal with highly noisy and missing input features due to its generative component. While these methods are promising attempts toward automating the joint AU intensity estimation, their current performance is still far below acceptable to replace human coders. For instance, [24] achieves average correlation rate (CORR) of 34.2% between the model prediction and human coders, on a subset of 5 AUs (1,2,3,4,6,9) from the DISFA dataset [20]. Furthermore, the highest CORR is achieved for AU1 (56.3%) and lowest for AU6 (11.9%). Likewise, [13] achieves average CORR of 43% on 12 AUs from DISFA dataset, with the best performance on AU25 (82%) and lowest on AU15 (11%). [15] achieves an intra-class correlation (ICC) score (a measure of raters agreement) of 77% on DISFA [20]. Yet, these experiments are performed in a subject-dependent manner, thus, are not representative for the target task. Nevertheless, note that for methods that perform subject-independent estimation of AU intensity, the CORR/ICC scores are far below what is acceptable for human coders. This is further affected by large variation in intensity estimation performance on individual AUs (being as low as 1% [13]).

While the methods mentioned above are still not accurate and reliable enough to replace the human coders, they can be exploited in order to reduce the burden of AU manual coding. This, in turn, would provide a better access to the labelled data that then can be used to improve models for AU intensity estimation. Several works have explored such approach. In Fast-FACS [4], FACS coders first detect AU peaks manually, and then an algorithm automatically detects their onset and offset phase. This led to more than 50% reduction in the time required for manual FACS coding. [27] developed an alternative approach that uses active learning. The system first performs initial labeling automatically. Then, a FACS coder manually makes any corrections if needed. The corrected labelling is then fed back to the system to re-train the model. In this way, system performance is iteratively improved. Likewise, [9] proposed an automatic method for successive detection of onsets, apexes, and offsets of consecutive facial expressions. All of these efforts

combine manual and automated methods with the aim of achieving synergistic increases in efficiency [3].

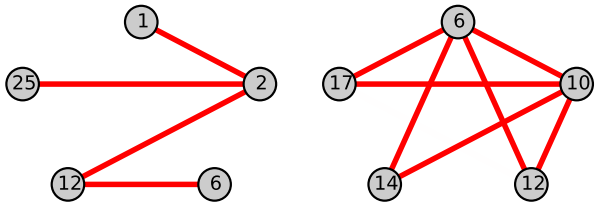
In this work, we propose a novel approach to guided coding of the AU intensity. This is in contrast to the works mentioned above that perform coding of the AU presence/absence, and do so independently for each AU. Specifically, we propose a coding framework that can reduce the burden of the manual coding by exploiting the newly introduced probabilistic model for joint estimation of the AU intensity, and the expert knowledge. This is attained efficiently in the following steps: the learned model for joint estimation of AU intensity 'assists' the FACS coders by providing estimates of the AU intensity level in the target face image. The model also provides uncertainty in these estimates, due to its probabilistic nature. This allows the FACS coders to reduce the range of possible choices for the target AU intensity by focusing on the most confident predictions by the model. The FACS coder then selects the correct intensity label for the target AU. This, in turn, is then used by the model to refine the intensity estimates for the remaining AUs, resulting in even more confident predictions by the model. This process is repeated until the intensities of all target AUs are scored. To this end, we introduce a novel Conditional Random Field (CRF) model for joint estimation of the AU intensity levels. This CRF model considers sparse, graph-induced, relationships between the intensity levels of multiple AUs simultaneously. We demonstrate the utility of the proposed approach on two benchmark datasets of spontaneous AUs, DISFA [20] and FERA2015 [26], in both model-driven and expert-driven settings (i.e., when the model predictions are refined using the coders' scoring, as depicted in Fig. 1). The contributions of this work can be summarized as follows:

1. We propose a novel framework for guided annotation of the AU intensity levels from face images. To the best of our knowledge, this is the first approach that allows so by combining the expert knowledge and the probabilistic modeling framework for automated joint estimation of the AU intensity.

2. We propose a novel CRF-like model for joint estimation of the AU intensity levels in real-time, making it suitable as an assistive tool for manual coding of AUs. Furthermore, this approach performs similarly or better than the state-of-the-art related approach for joint inference of AU intensity [24], while being computationally more efficient.

3. We show on two public datasets of face images, coded in terms of the AU intensity levels, that the proposed approach achieves significantly better estimation of the AU intensity levels (due to the joint AU modeling) when more AU labels are provided by the expert. This can be exploited to largely speed-up the manual FACS's coding efforts by providing the label suggestions to the coders, along with their uncertainty.

4. We also show that comparable gains in the estimation performance can be obtained by the proposed approach regardless of the labelling order (from the most to the least uncertain AUs by the model, and vice verse). This, in turn, allows the FACS coder to start annotating the 'easiest' AUs first, resulting in more confident predictions for more difficult AUs, thus, reducing the labelling effort.



(a) DISFA (b) FERA2015

Figure 2: The glasso removes the majority of AU pairs from the precision matrix, preserving only the strongest partial correlations. These are later modeled in the proposed CRF.

In the rest of the paper, we introduce the proposed model for joint modeling of AU intensity levels, and describe our approach for combining the model learning and expert knowledge in order to perform guided coding of the AU intensity. We then perform experimental evaluation of the proposed approach, and conclude the paper.

2. Methodology

2.1. CRF for Joint AU Intensity Estimation

Let us denote the training set as $\mathcal{D} = \{\mathbf{Y}, \mathbf{X}\}$. $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_i, \ldots, \mathbf{y}_N]^T$ is comprised of N instances of multivariate outputs stored in $\mathbf{y}_i = \{\mathbf{y}_i^1, \ldots \mathbf{y}_i^q, \ldots \mathbf{y}_i^Q\}$, where Q is the number of AUs, and $\mathbf{y}_i^q$ takes one of $\{1, ..., L^q\}$ discrete intensity levels of the q-th AU. Furthermore, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_N]^T$ are input features (e.g., facial points) that correspond to the combinations of labels in $\mathbf{Y}$. Thus, our goal is to simultaneously estimate the combination of the intensity levels $\mathbf{y}^q$ of Q AUs, given the facial features $\mathbf{x}$. To this end, we propose a CRF structure-based approach. First, we build a dependency graph which takes the form $T = (V, E)$, where the vertices $V = v_1, ..., v_Q$ are used to model each AU, and edges ($e_{rs} \in E$) account for dependencies between AU intensities.

2.2. The Graph Structure

Modeling the fully connected graph (i.e., $Q \times (Q-1)/2$ edges) is impractical as not all AU exhibit a dependence pattern (e.g., AU16 (lower lip depressor) and AU17 (chin raiser) rarely co-occur). We learn the cliques (i.e., the edges) in our CRF model from the precision matrix derived from the correlation matrix S of the intensity labels of AUs. This is because the precision matrix unravels partial correlations among the AUs, while the correlation matrix focuses on marginal correlations [10]. Important advantage of this is that AUs correlated through another AU are ignored, therefore, avoiding a redundant modeling. To this end, we exploit partial correlations using a sparse estimate of the precision matrix Υ computed from S. The aim is to reduce the number of the model parameters by not accounting for 'weak' dependencies among the AUs. We first empirically estimate S from training data, and then obtain sparse $\tilde{S}$ by means of the graphical lasso estimation [7], used to solve the following convex optimization:

$$(\Upsilon, \tilde{S}) = \min_{\Upsilon \succ 0} -\ln\det(\Upsilon) + tr(S\Upsilon) + \kappa\|\Upsilon\|_1, \qquad (1)$$

where κ is the regularization parameter.[1] Finally, the edge set E is defined by keeping the edges that satisfy: $E = \{(r,s) : |\Upsilon_{r,s}| > \delta\}$. $\delta = 0.05$ is chosen so that only the pairs of AUs with strong partial correlations are kept, resulting in a model with fewer parameters [21]. The learned graphs for the used datasets are depicted in Fig. 2.

2.3. Structured Learning

Using CRFs [14], the joint pdf of Q random variables (AU intensity) is defined as:

$$P(\vec{y}|\vec{x}, \Omega) = \frac{1}{Z}\prod_c \Psi(\mathbf{y}_c|x) \qquad (2)$$

where Z is the partition function, $\mathbf{y}_c$ is the subset of random variables in clique c, $\Psi(\cdot)$ is the edge potential defined on the labels in this clique, as explained below, and $\Omega = \{\vartheta, \theta\}$ are the model parameters.[2] In our setting, we only consider unary and binary cliques, modeling individual independent AUs and pairs of AUs. In other words, $\mathcal{C} = V \cup E$, where E is the set of edges in $\mathcal{G}$. Hence,

$$\Psi(\mathbf{y}_c|x) = \begin{cases} \exp(f_n(y_r, \vec{x})), & \begin{matrix} c = r \in V \\ \text{unary clique} \end{matrix} \\ \exp(f_{rs}(y_r, y_s, \vec{x})), & \begin{matrix} c = (r,s) \in E \\ \text{pairwise clique} \end{matrix} \end{cases} \qquad (3)$$

$f_r(y_r)$ stands for the node features that correspond to AU_r with intensity level y_r. Similarly, $f_{rs}(y_r, y_s)$ describes the edge features which correspond to the compatibility of AU_r and AU_s with the intensities y_r and y_s. The choice of the node $f_r(y_r)$ and edge $f_{r,s}(y_r, y_s)$ features depends on the target task, and plays a crucial role in the definition of CRFs. Furthermore, we assign linear weights to the node features as:

$$f_r(y_r, \vec{x}) = \sum_k^L I(y_r = k) \cdot \vec{w}_{rk}^T \cdot \vec{x} + b_{rk} \qquad (4)$$

where L is the number of intensity levels and $I(*)$ is the indicator function that returns 1 (0) if the argument is true (false). The projection vector $\vec{w}_{rk}^T$ performs feature selection for the AU scores as y_r, and b_{rk} is a bias for that AU. The edge features model the dependence between two AUs as:

$$f_{rs}(y_r, y_s, \vec{x}) = \sum_{m,k}^L I(y_r = m, y_s = k) \cdot u_{rs[m,k]} \qquad (5)$$

where $m, k = 1, ..., L$ are combinations of intensity levels and u_{rs} measures the dependence between each AU intensity combination.

2.4. Optimization

By using the notion of the negative log-likelihood, our learning objective can be written as:

$$NCL = -\sum_{i=1}^N \{\log(\Psi(\mathbf{y}_i|x)) - \log(Z)\}, \qquad (6)$$

where N is the number of training instances. The most critical aspect in evaluation of the joint distribution in Eq. 2 is computation of the partition function Z. This is an np-complete problem, and thus, exact inference is intractable in general. However, approximate methods based on Markov chain Monte Carlo (MCMC) and Loopy Belief Propagation (LBP) have been proposed to this aim [28]. To this end, we resort to the message-passing LBP, which is a dynamic programming approach to approximating conditional probability queries in a graphical model. For completeness, we briefly summarize the LBP algorithm: each node y represents one AU and computes a belief $BEL(y_q = i) = P(y_q = i|ev)$, where ev is the observed evidence to intensity level i. This is attained by combining messages from its adjacent nodes. This procedure is repeated for each node until convergence. The running time for the LBP algorithm on our graph is $\mathcal{O}(Q * L^C)$, where Q is the number of AUs, L is the number of intensity levels, and C is the maximum clique size [1]. The convergence was reached in less than 30 iteration and the total processing is done in real time (e.g., 12K frames on an 2.4 GHz Intel Xeon CPU processed in less than 25 sec). Lastly, the parameter optimization is performed by approximating the partition function and minimizing the NCL (Eq.6) w.r.t. Ω. For this, we employ LBP and the Conjugate gradient method with line search [22].

[1] We used the glasso Matlab code from [7].

[2] For simplicity, we often drop the dependency on Ω in notations.

Algorithm 1 Inference: $\text{CRF}^{+}_{\text{exp}}$

Input: Test data: $\mathcal{D} = \{\vec{x}_i\}_{i=1}^{N}$
Model parameters: $\Omega = \{w, u\}$
Graph Structure: $T = \{V, E\}$
Output: Annotations: Y

Initialize output and potentials
Y={}
$\forall r \in V \rightarrow n_r = f_r(\vec{x})$ (Eq. 4)
$\forall rs \in E \rightarrow e_{rs} = f_{rs}(\vec{x})$ (Eq. 5)
repeat
Step-1: compute marginals for all AUs (Eq. 8)
$\forall r \in V \rightarrow p_r = P(r|Y)$
Step-2: select AU for annotation (Eq. 11)
$i \leftarrow \underset{r}{\operatorname{argmin}}\, H(p_{[r]})$
Step-3: add labels to observed set
$Y \leftarrow Y \wedge y_i$
until all AUs are labelled

Regularization: During training, we seek to find parameters Ω^* by solving the regularized optimization problem:

$$\Omega^* = \arg\min_{\Omega^*} NCL(\varphi, \theta) + R_n + R_e,$$

where NCL is defined by Eq.6, R_n and R_e stand for the standard L_2 regularization of the model parameters $\vec{w}_r$ and u_{rs}, respectively.

2.5. Inference

In this section, we briefly describe the proposed approximate inference method based on an iterative maximum a-posterior (MAP) algorithm. First, we perform MAP inference using the LBP for the CRFs to obtain the initial estimates of the AU intensity. In the following iterations, we compute the marginal node probability, conditioned on a set of observed AUs as provided by FACS coders. This results in two sets of nodes, the set of observed (Y) and predicted (B) nodes. The conditional probability of B given Y is then defined by:

$$P(B|Y) = P(Y, B)/P(Y) \tag{7}$$

The marginal probability for the i-th AU having intensity k, given the labels Y can then be computed by marginalizing out the predictions of all remaining AUs from B.

$$P(B_i = k|Y) = \sum_{B \neq B_i} P(B|Y) \tag{8}$$

For clarity, we use an example with 3 AUs. Assuming we have a trained model that returns the joint predictions for all three AUs. We also have a dataset in which the labels of the third AU are given $Y = \{y_3\}$. Using Eq. 7, the conditional probability of $B = \{y_1, y_2\}$ can be directly computed by:

$$P(y_1, y_2|y_3) = \frac{P(y_1, y_2, y_3)}{\sum_{y_1, y_2} P(y_1, y_2, y_3)} \tag{9}$$

Where $P(y_1, y_2, y_3)$ is defined in Eq. 2. The marginal probability of node y_1 having intensity level k can be directly computed by:

$$P(y_1 = k|y_3) = \sum_{y_2} P(y_1 = k, y_2|y_3) \tag{10}$$

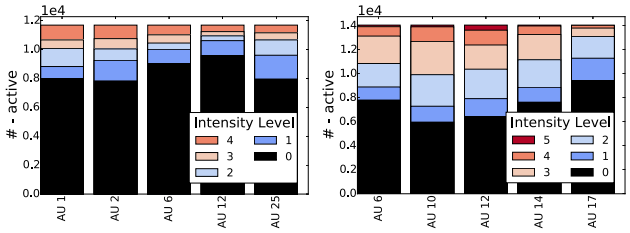
The MAP solution and the marginals can be computed by iteration over all possible combinations of intensity levels for y_2. For larger problems however, this is again intractable and we estimate the values using LBP (see above).

2.6. Expert Driven Entropy Reduction

The more outputs to estimate, the more uncertainty in the estimation process. We propose an estimation approach applicable to any probabilistic multi-output classifier, aiming at obtaining more accurate predictions by providing additional partially labeled data. This is done by iteratively adding evidence to the predicted outputs in order to minimize the uncertainty for the remaining ones. In our CRF model, this is equivalent to minimizing the entropy of the predicted probability distribution. Entropy is a measure of the expected information content or uncertainty of a probability distributions and is defined as:

$$H_r = -\sum_k P_{rk} \cdot log(P_{rk}), \tag{11}$$

where P_{rk} is the probability of AU_r having the intensity level k. This reflects the notion that the lower the probability of an intensity level to occur, the higher amount of information in the message stating that the level occurred. In this work, we aim to increase the classification performance for joint AU intensity estimation by providing labels of certain AUs in an iterative manner. This can be done in two different settings either by manually annotating AUs, that are difficult to predict (high entropy) or by annotating AUs that are easy to predict (low entropy). For example, a joint model could be trained for predicting AU-intensities. A FACS coder could then simply annotate the easy targets (e.g., intensity of AU12) and some of the remaining AU intensity predictions should benefit from this additional information. In what follows, we apply iterative labeling, starting from the AU with the lowest entropy and recompute the performance on the remaining AUs. A trained model for joint prediction, a graph that defines the dependency structure and a set of input vectors is used as input. Moreover, we initialize the output Y with an empty list and compute the node and edge potentials. In the first iteration, we compute



(a) DISFA (b) FERA2015

Figure 3: Distribution of the AU intensity levels.

the marginals (Step-1) and the entropy (Step-2) for all AUs. The expert annotator assigns a score to the AU with the lowest entropy (Step-3). Finally, we include the label of that AU to the output set Y and recompute the predictions (Step-1) with a lower entropy (i.e., higher certainty and, thus, expected higher classification performance). These 3 steps are repeated until all AUs are annotated, as described in Alg. 1.

3. Experiments

3.1. Datasets

We evaluate the proposed approach on two benchmark datasets - Denver Intensity of Spontaneous Facial Actions (DISFA) [20], and on a subset of the BinghamtonPittsburgh 4D Spontaneous Expression (BP4D) [26] database that is a part of the FERA2015 challenge for the AU-intensity estimation. This databases include acted and spontaneous expressions and vary in image quality, video length, annotation, number of subjects, and context. Specifically, the **DISFA** dataset contains videos of 27 subjects watching YouTube videos. For the experiments presented here, we used a subset of 5 AUs - two pairs of of highly correlated AUs (1,2 and 6,12) and the most frequently occurring AU (25). Only the image frames with at least two active AUs (intensity levels > 1) were used to balance the data. To this end, we further merged levels 5 and 6 as only few examples of the highest intensity levels were present. The **FERA2015** database includes video of 41 subjects: 21 in the training, and 20 in the development (in our case, test) set. The dataset contains intensity annotations for AUs 6, 10, 12, 14, and 17. The resulting intensity distribution is depicted in Fig. 3.

3.2. Features

We used the geometric facial features in our experiments, as in [13]. Namely, we used the locations of 49 out of 66 fiducial facial points (provided by the database creators) extracted from facial images in each dataset, using the 2D Active Appearance Model (2D-AAM) [19]. We removed the points from the chin line, as these do not affect the estimation of target AUs. We then registered the 49 facial points to a reference face (average points in each dataset) using an affine transformation. To reduce the dimensionality of the features, we applied PCA, retaining 97% of the energy. This resulted in approximately 20 dimensional feature vectors.

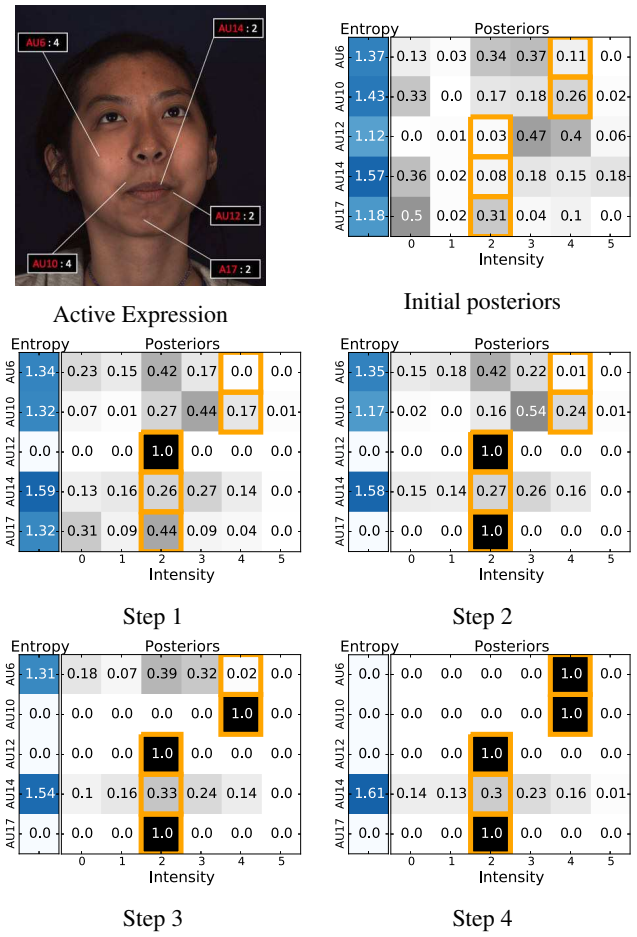


Figure 4: The guided coding of AU intensities in an example facial expression. The posteriors provide the probability of each AU intensity by the model. The yellow squares depict the correct labels. In each step, the FACS coder labels the AU with the least average entropy (bar on the left), and then the model predictions are updated. This is repeated until all AU are coded. Observe the decrease in the average entropy of each AU, and the model automatically 'correcting' its predictions after each step.

3.3. Evaluation metrics

Intra-class Correlation (ICC). We report the Intra-class Correlation (ICC(3,1)) [25], which is commonly used in behavioral sciences to measure agreement between annotators (in our case, the AU intensity labels and model predictions).

	Database	DISFA						FERA2015					
	AU	AU1	AU2	AU6	AU12	AU25	avg.	AU6	AU10	AU12	AU14	AU17	avg.
ICC	MLR	.23	.39	.67	.35	.60	.45	.35	.48	.81	.22	.27	.43
	CRF	.40	.52	**.73**	.52	.63	.56	.49	.48	.71	.23	.23	.43
	MRF[24]	.42	.55	.69	.54	.54	.55	.53	.44	.83	.23	.24	.45
	CRF_{auto}	.41	.56	**.73**	.51	.61	.56	.49	.46	.73	.21	.23	.42
	MRF_{auto}	.38	.44	.69	.58	.51	.52	.53	.42	.79	.25	.21	.44
	CRF^{-}_{exp}	**.83**(3)	.52(1)	**.76**(5)	.53(4)	.63(2)	.65	.55(4)	.48(1)	.82(5)	.24(2)	.28(3)	.47
	MRF^{-}_{exp}	.73(5)	.76(4)	.64(2)	**.76**(3)	.54(1)	**.69**	**.56**(4)	.46(2)	**.87**(5)	.23(1)	.30(3)	.48
	CRF^{+}_{exp}	.46(4)	**.84**(5)	**.73**(1)	.71(2)	**.69**(3)	**.69**	.49(1)	**.63**(4)	.83(2)	.33(5)	**.31**(3)	**.52**
	MRF^{+}_{exp}	.42(1)	.74(2)	.66(3)	**.76**(4)	.60(5)	.64	.53(1)	.62(4)	**.87**(2)	**.36**(3)	.28(4)	**.53**
F1	MLR	.33	.22	.48	.19	.21	.29	.23	.23	.40	.18	.18	.24
	CRF	.36	.29	**.50**	.22	.29	.33	.29	.33	.35	.25	.20	.28
	MRF[24]	.32	.31	.43	.24	.32	.32	.32	.32	.43	.25	.21	.31
	CRF_{auto}	.36	.28	.49	.21	.30	.33	.29	.31	.35	.22	.21	.28
	MRF_{auto}	.32	.24	.41	.21	.27	.29	.32	.29	.41	.24	.23	.30
	CRF^{-}_{exp}	.39(3)	.29(1)	.37(5)	.29(4)	**.50**(2)	.37	.30(4)	.33(1)	.44(5)	.25(2)	.22(3)	.31
	MRF^{-}_{exp}	**.41**(5)	.32(4)	.45(2)	.39(3)	.32(1)	.38	**.34**(4)	.31(2)	**.51**(5)	.25(1)	.24(3)	.33
	CRF^{+}_{exp}	.36(4)	.39(5)	**.50**(1)	.23(2)	.49(3)	**.39**	.29(1)	**.44**(4)	.42(2)	.31(5)	**.25**(3)	.34
	MRF^{+}_{exp}	.32(1)	**.51**(2)	.32(3)	**.40**(4)	.45(5)	**.40**	.32(1)	.42(4)	.47(2)	**.34**(3)	.24(4)	**.36**

Table 1: Performance on the DISFA and FERA2015 datasets. The iterative methods with expert labeling perform significantly better than classical models (MLR,CRF,MRF). The iteration order is shown in brackets. The best results are depicted in bold.

F1 Score. We also report the F1 score which is widely used for facial expression recognition tasks because of its robustness to the imbalanced labels (see Fig. 3). For each AU intensity, F1 is computed based on a frame-based intensity scores and by averaging the scores for all levels.

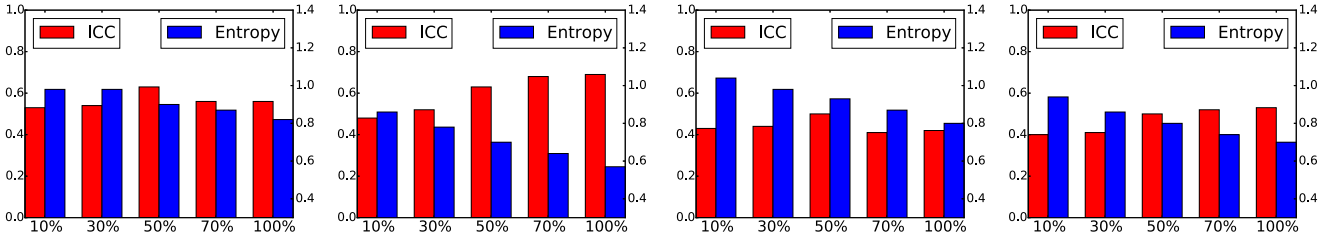
3.4. Models

We evaluate the performance of the proposed annotation framework in different settings using the proposed CRF. We also provide comparisons with the related state-of-the-art MRF model [24] for joint modeling of AUs. For the **MRF**, we compute the map solution of each minimum spanning tree of the target AU separately, as done in [24]. As a baseline, we include multivariate logistic regression **MLR**, derived by removing the edge potentials from our CRF. $\mathbf{CRF_{auto}}$ and $\mathbf{MRF_{auto}}$ represent the fully automatic setting in which labels are provided using solely the model predictions. The predictions for each AU are obtained by computing the MAP solution conditioned on the labels that have already been obtained in the previous steps by the model. This is applied iteratively, starting with the AU with the lowest entropy. Similarly, $\mathbf{CRF^{+}_{exp}}$ and $\mathbf{MRF^{+}_{exp}}$, also apply iterative computation of the MAP solution for each AU and select the one with the lowest entropy for annotation. The difference to the automatic setting is that here, in each iteration[3], we assign the ground truth label to the target AU, as provided by the FACS coder. The same is performed in reverse order (decreasing order starting with the highest entropy) in $\mathbf{CRF^{-}_{exp}}$ and $\mathbf{MRF^{-}_{exp}}$ (see Alg.1). For DISFA, we performed a 3 fold subject-independent cross-validation, while for FERA2015 development set is used for testing.

[3]In each iteration, one AU is coded by an expert. Thus, the number of iterations is the number of AUs - 1.

3.5. Evaluation

The results on the DISFA and FERA2015 databases are shown in Table 1. Note that to report the average results for the expert-driven models ($\mathbf{CRF^{-}_{exp}}$ and $\mathbf{MRF^{-}_{exp}}$), the order of the AUs (given in the brackets) is determined correspondingly based on the average entropy for each AU, deduced from the training set. The expert-driven models outperform the fully automatic models on both databases and in both measures, F1 and ICC, as expected. In particular, we make two observations: first, this improvement is most significant in the increasing entropy expert annotated models. In this configuration, the last AU is the most difficult to predict and it benefits most from the expert scoring. Secondly, the classifiers are much more certain about the decision of one AU if it is highly correlated to an already annotated AU, as is the case for AU6 and AU12 in FERA2015, and AU1 and AU2 in DISFA. In this case, the ICC increased 10% for AU12 when the labels of AU6 are given, and 32% for AU2 if the labels of AU1 were given, respectively. This is also expected, since these AU pairs are highly correlated. Furthermore, CRF^{+}_{exp}outperforms MRF^{+}_{exp}on the DISFA database because its dependency structure is obtained from the graphical lasso and can contain loops that allow it to model more complex relationships among groups of AUs. However, this is less pronounced in FERA2015, where both models achieve similar results. A reason for this is that in FERA2015 some frames are not fully annotated (at least one AU label is missing in about 40% of the frames), which may result in false



(a) DISFA: CRF (b) DISFA: CRF^{+}_{exp} (c) FERA2015: CRF (d) FERA2015: CRF^{+}_{exp}

Figure 5: Performance and entropy for different subsets of the DISFA and FERA2015 database. The average ICC and entropy is computed on that portion of frames with the highest entropy.

correlations learned by our model.

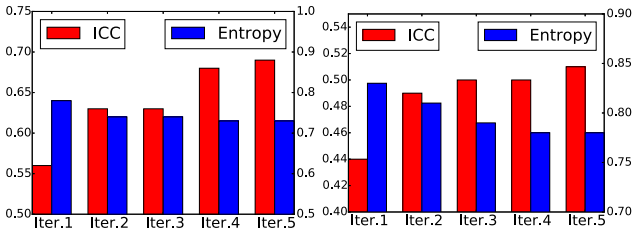
Fig. 4 shows the predictions for each AU intensity level in an expressive face from the FERA2015 database. This sample sample shows a facial expression in which the intensity levels are not obvious to identify. Note that in the initial step, the CRF model would return a faulty intensity prediction for AU14. However, if the expert coder assigns the labels for AU12 (step 1) and AU17 (step 2), the resulting marginals of AU14, conditioned on the given intensity levels, are changing and predicting the right value. This can be seen in step 3, where the probability mass is distributed around the correct score for AU14.

Fig. 5 shows the average performance and entropy on a subset of the 10%, 30%, ..., 100% frames with the highest entropy from the test set. As expected, the ICC has the lowest value on both databases if using only 10% of frames with the highest entropy (highest uncertainty) and the ICC is increasing when adding frames with the lower entropy. With the CRF^{+}_{exp}model, the increase in performance is more pronounced on DISFA, also observed from Table 1. Note that the CRF model reaches the highest performance with 70% of samples on the DISFA, and 50% of samples on the FERA2015 database. The performance drops if adding additional samples with a lower entropy. This shows that there is a portion of the frames for which the CRF model is falsely overconfident. However, that effect is not observed in the CRF^{+}_{exp}models, boosted by the expert knowledge.

Finally, Fig. 6 shows the average entropy decrease and performance increase with each iteration if the FACS coder does guided coding (Alg.1). The performance improves significantly within the first 3 iterations (i.e., annotation of 3 AUs). The performance on the remaining AUs clearly increases and the predictions can be used as a hint for scoring remaining AUs by looking into their average entropy and the probability estimates for each level.

4. Conclusions

We proposed an expert-driven probabilistic approach for joint modeling and estimation of AU intensities. We showed



(a) DISFA (b) FERA2015

Figure 6: Entropy and performance per iteration. The plots show the entropy decrease per iteration of the CRF^{+}_{exp}model. As expected, ICC increases with each step.

that the classification performance can significantly be increased when partial label information is provided by the FACS coder. Specifically, we showed that the AU coding burden can be reduced by means of automated estimation of the AU intensity, which becomes significantly more accurate by providing the AU scores to the model in the iterative fashion. This, in turn, allows the FACS coder to faster make decision about the target AU intensity by observing the estimated entropy of each AU. In our future work, we plan to extend this approach by also updating the classifier after each label is provided. Hopefully, this will lead to an even more accurate model predictions, and, consequently, a more efficient AU coding process.

Acknowledgments

This work is funded by the European Community Horizon 2020 under grant agreement no. 645094 (SEWA) and no. 688835 (DE-ENIGMA). The work of V. Pavlovic is funded by the National Science Foundation under Grant no. IIS0916812. The work of J. Cohn is supported in part by the National Institute of Mental Health of the National Institutes of Health under Award Number MH096951 to the University of Pittsburgh. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] C. M. Bishop. *Pattern recognition and machine learning*. 2006.

[2] W.-S. Chu, F. D. L. Torre, and J. F. Cohn. Selective transfer machine for personalized facial action unit detection. *CVPR*, pages 3515–3522, 2013.

[3] J. F. Cohn and F. De la Torre. Automated face analysis for affective computing. *The Oxford handbook of affective computing*, page 131, 2014.

[4] F. De la Torre, T. Simon, Z. Ambadar, and J. F. Cohn. Fast-facs: A computer-assisted system to increase speed and reliability of manual facs coding. pages 57–66. 2011.

[5] P. Ekman, W. V. Friesen, and J. C. Hager. Facial action coding system. *Manual: A Human Face*, 2002.

[6] S. Eleftheriadis, O. Rudovic, and M. Pantic. Multi-conditional latent variable model for joint facial action unit detection. *ICCV*, 2015.

[7] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, pages 432–441, 2008.

[8] J. M. Girard, J. F. Cohn, and F. De la Torre. Estimating smile intensity: A better way. *Pattern recognition letters*, 66:13–21, 2015.

[9] Z. Hammal. Efficient detection of consecutive facial expression apices using biologically based log-normal filters. *Advances in Visual Computing*, pages 586–595, 2011.

[10] S. Horvath. *Weighted Network Analysis: Applications in Genomics and Systems Biology*. Springer Science & Business Media, 2011.

[11] L. A. Jeni, J. M. Girard, J. F. Cohn, and F. D. L. Torre. Continuous au intensity estimation using localized, sparse facial feature space. *FG*, pages 1–7, 2013.

[12] S. Kaltwang, O. Rudovic, and M. Pantic. Continuous pain intensity estimation from facial expressions. *Advances in Visual Computing*, pages 368–377, 2012.

[13] S. Kaltwang, S. Todorovic, and M. Pantic. Latent trees for estimating intensity of facial action units. *CVPR*, 2015.

[14] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *ICML*, pages 282–289, 2001.

[15] Y. Li, S. M. Mavadati, M. H. Mahoor, and Q. Ji. A unified probabilistic framework for measuring the intensity of spontaneous facial action units. *FG*, 2013.

[16] P. Lucey, J. F. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett, and K. M. Prkachin. Automatically detecting pain in video through facial action units. *TSMCB*, pages 664–674, 2011.

[17] M. Mahoor, S. Cadavid, D. Messinger, and J. Cohn. A framework for automated measurement of the intensity of non-posed facial action units. *CVPR*, pages 74–80, 2009.

[18] M. H. Mahoor, M. Zhou, K. L. Veon, S. M. Mavadati, and J. F. Cohn. Facial action unit recognition with sparse representation. *FG*, pages 336–342, 2011.

[19] I. Matthews and S. Baker. Active appearance models revisited. *IJCV*, pages 135–164, 2004.

[20] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. Disfa: A spontaneous facial action intensity database. *TAC*, pages 151–160, 2013.

[21] R. Mazumder and T. Hastie. Exact covariance thresholding into connected components for large-scale graphical lasso. *JMLR*, pages 781–794, 2012.

[22] C. Rasmussen and C. Williams. *Gaussian processes for machine learning*. The MIT Press, 2006.

[23] O. Rudovic, V. Pavlovic, and M. Pantic. Context-sensitive dynamic ordinal regression for intensity estimation of facial action units. *TPAMI*, pages 944–958, 2014.

[24] G. Sandbach, S. Zafeiriou, and M. Pantic. Markov random field structures for facial action unit intensity estimation. *ICCV*, 2013.

[25] P. E. Shrout and J. L. Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.

[26] M. F. Valstar, T. Almaev, J. M. Girard, G. McKeown, M. Mehu, L. Yin, M. Pantic, and J. F. Cohn. Fera 2015-second facial expression recognition and analysis challenge. *FG*, 6:1–8, 2015.

[27] L. Zhang, Y. Tong, and Q. Ji. Active image labeling and its application to facial action labeling. *ECCV*, pages 706–719, 2008.

[28] Y. Zhang and J. Schneider. A composite likelihood view for multi-label classification. *Artificial Intelligence and Statistics*, pages 1407–1415, 2012.

[29] K. Zhao, W.-S. Chu, F. De la Torre, J. F. Cohn, and H. Zhang. Joint patch and multi-label learning for facial action unit detection. *CVPR*, 2015.