Joint Representation and Estimator Learning for Facial Action Unit Intensity Estimation

Yong Zhang[1], Baoyuan Wu[1]*, Weiming Dong[2], Zhifeng Li[1], Wei Liu[1], Bao-Gang Hu[2], and Qiang Ji[3]
[1]Tencent AI Lab, [2]National Laboratory of Pattern Recognition, CASIA, [3]Rensselaer Polytechnic Institute
{zha6.5ngyong201303,wubaoyuan1987}@gmail.com, weiming.dong@ia.ac.cn
michaelzfli@tencent.com, wl2223@columbia.edu, hubg@nlpr.ia.ac.cn, qji@ecse.rpi.edu

Abstract

Facial action unit (AU) intensity is an index to characterize human expressions. Accurate AU intensity estimation depends on three major elements: image representation, intensity estimator, and supervisory information. Most existing methods learn intensity estimator with fixed image representation, and rely on the availability of fully annotated supervisory information. In this paper, a novel general framework for AU intensity estimation is presented, which differs from traditional estimation methods in two aspects. First, rather than keeping image representation fixed, it simultaneously learns representation and intensity estimator to achieve an optimal solution. Second, it allows incorporating weak supervisory training signal from human knowledge (e.g. feature smoothness, label smoothness, label ranking, and positive label), which makes our model trainable even fully annotated information is not available. More specifically, human knowledge is represented as either soft or hard constraints which are encoded as regularization terms or equality/inequality constraints, respectively. On top of our novel framework, we additionally propose an efficient algorithm for optimization based on Alternating Direction Method of Multipliers (ADMM). Evaluations on two benchmark databases show that our method outperforms competing methods under different ratios of AU intensity annotations, especially for small ratios.

1. Introduction

Facial Action Coding System (FACS) [7] defines AUs to depict facial muscle movements. It quantifies AU intensities into 6 ordinal levels and provides instructions for annotation. AU intensity estimation is challenging partially due to the lack of annotations. Reliable AU intensity labels can be annotated by trained human coders. It is convenient to capture a large set of expression sequences through digital cameras. However, it requires great effort to annotate every frame, which makes it difficult to construct a large database.

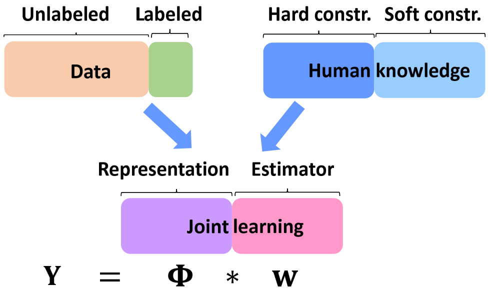

Figure 1: The diagram of joint learning for AU intensity estimation. It simultaneously learns image representation and intensity estimator with limited annotations. Human knowledge is incorporated as hard or soft constraints to provide weak supervision.

The performance of AU intensity estimation is determined by several factors, including image representation, intensity estimator, and supervisory information. Most existing methods [16, 10] focus on estimator learning, regardless of image representation and unlabeled frames. Representation is firstly learned by an unsupervised or supervised method. Then, the estimator is trained in the pre-learned feature space. However, the pre-learned feature space is not guaranteed to fit well the estimator learning since they are learned separately without considering their connections. Recently, deep learning provides an end-to-end strategy to learn a mapping from the input to the output [8]. It can be treated as a joint representation and estimator learning method. But deep models require a large amount of labeled samples to avoid overfitting. Though [41] learns a weakly supervised deep model, it still requires thousands of images.

Besides, few works focus on intensity estimation with limited annotations except for [26, 43, 42, 41]. Human knowledge such as label smoothness or label ranking are exploited to compensate the lack of annotations. [26] encodes label smoothness through the structure of dynamic models while [42] emphasizes the smoothness on relevance

[*] Corresponding author.

through regularization. Label ranking has been studied in ordinal regression [9] and applied to intensity estimation in [43, 42, 41]. Other types of general knowledge such as feature smoothness and positive intensity are rarely studied. The smoothness refers to soft constraint while label ranking and positive intensity refer to hard constraint. Though there exist various types of human knowledge, existing methods always leverage one of them and there is no general framework to incorporate different types of knowledge.

Joint representation and estimator learning and human knowledge have been studied individually, but there lacks a framework to simultaneously incorporate all of them. To alleviate this issue, we propose a general framework for AU intensity estimation (see Fig. 1), which can not only jointly learn representation and estimator with weakly labeled sequences, but also flexibly incorporate different types of knowledge on AU intensity and image representation. *Weak annotation* refers that locations of peak and valley frames are firstly identified in training sequences (*qualitative annotation*) and then intensities of few selected frames are labeled (*quantitative annotation*). Identification of peak and valley frames can be performed according to their definitions in [17], which is much easier to obtain than quantitatively annotating every frame. The proposed framework has several advantages. Firstly, representation and estimator are jointly optimized to make accurate intensity prediction. The joint learning can obtain a better solution than individual learning. Secondly, human knowledge involves both representation and estimator. It can provide weak supervision for joint learning and make it feasible to exploit unlabeled frames more efficiently by using relationships among labeled and unlabeled frames. Thirdly, this framework requires scarce intensity annotations to obtain satisfied performance, while deep models need much more.

Our main contributions are three-fold. **(1)** We propose a general framework for AU intensity estimation to jointly learn representation and estimator with very limited intensity annotations and to encode different types of human knowledge. **(2)** We develop an efficient algorithm based on the ADMM framework to optimize the formulated problem. **(3)** Evaluations under different annotation ratios are performed on two benchmark databases to demonstrate the superior performance of the proposed method.

2. Related Work

AU intensity estimation. Most existing approaches use supervised techniques for AU intensity estimation such as [16, 10, 38, 22, 32, 21, 12, 36]. They leverage only labeled images. Several methods exploit spatial relationships among the intensities of multiple AUs through probabilistic graphical models, such as tree-structured Markov random field [27], copula conditional random field [35]. Kaltwang *et al.* [11] propose a latent tree model by learning graph structure from features and labels. The co-occurrence of AU pairs is spatial label smoothness which is implicitly encoded in potential functions. Besides, several works consider temporal label smoothness by using dynamic models, such as hidden Markov model [18], dynamic Bayesian networks [14], context-sensitive conditional ordinal random field [24]. Temporal label smoothness is implicitly encoded in the model structure. However, these supervised models require frame-level intensity annotations and they only focus on estimator learning, regardless of representation learning. Deep models have made astonishing progress in different fields due to their large model complexity. Supervised deep models have been used to estimate AU intensity [8, 34, 31]. However, they contain millions of parameters and require a large amount of annotated images for training.

Few works focus on using unlabeled images for AU intensity estimation. Multi-instance learning (MIL) has been used for event detection such as MS-MIL [29], RMC-MIL [25], and LOMo [30]. Event detection is a binary classification problem, but the AU intensity has multiple levels. They can not be directly applied. Only methods [26, 43, 42, 41] exploit unlabeled images for AU intensity estimation. Ruiz *et al.* [26] propose Multi-instance Dynamic Ordinal Random Fields (MI-DORFs) by exploiting multi-instance learning to treat each sequence as a bag. Zhang *et al.* [42] propose a bilateral ordinal multi-instance regression model (BORMIR). Both use temporal label smoothness to exploit unlabeled images. Zhao *et al.* [43] estimate expression intensity estimation by combining ordinal regression and SVR (OSVR). They use label ranking to exploit unlabeled images. These methods learn only intensity estimator or leverage only one type of domain knowledge, regardless of image representation. Differently, our method jointly learns estimator and representation. We simultaneously incorporate all types of knowledge such as features smoothness, label smoothness, label ranking, and positive intensity. Besides, knowledge is applied in the subspace rather than in the original feature space. A weakly supervised deep model is used in [41], but it still requires thousands of labeled images. Differently, our method performs the joint learning with only few hundreds of labeled images and is applicable to a small database.

Semi-supervised learning. Semi-supervised learning methods learn models with both labeled and unlabeled images. Different assumptions are made on the correlation between sample representation and target label, including smoothness assumption [15], cluster assumption [4], and manifold assumption [1, 20]. Kim *et al.* [13] consider the second-order Hessian energy for semi-supervised regression (HSSR) under the manifold assumption. These methods leverage unlabeled images, but they do not learn the representation. Zhang *et al.* [40] proposed a convex subspace learning (CSL) approach by combining unsupervised subspace learning and supervised classifier learning. Compared to [40], our method incorporates various types of human knowledge to leverage unlabeled images more efficiently.

3. The Proposed Approach

3.1. Weak annotation

The smoothness of muscle movements leads to the smooth evolution of facial appearance. AU intensity also evolves smoothly in sequences if the frame rate of camera is high enough to capture subtle changes of facial appearance. Weak annotation consists of two parts, *i.e.,* qualitative annotation and quantitative annotation. Qualitative annotation refers to identifying the locations of key frames, *i.e.*, peak and valley frames. Quantitative annotation refers to annotating AU intensities of a small set of frames in sequences. Though multiple peaks and valleys exist in a sequence, they occupy a small portion in the whole database. Weak annotation contains the locations of key frames and intensity annotations of few frames, which is easier to achieve than labeling every frame. Given the weak annotation, training sequences can be split into segments according to the locations of key frames. AU intensity monotonically evolves in each training segment and it has three types, *i.e.*, increasing, decreasing, and keeping the same. Following [42], to avoid using an extra variable to specify the trend, we reverse the frame ordering of segments that evolve from a peak to a valley. Finally, AU intensity increases or keeps the same in all training segments.

3.2. Problem statement

Given unlabeled expression sequences, the locations of key frames are firstly identified. Then, intensities of partial randomly selected frames are labeled. The training set consists of two parts. One part is qualitatively labeled segments, $\mathcal{D}_s = \{\mathbf{S}_u^m\}_{m=1}^M$, where $\mathbf{S}_u^m = \{S_u^{m,t}\}_{t=1}^{T_m} \in \mathbb{R}^{T_m \times d}$ denotes the features of frames in the m-th segment and T_m is the number of frames. $\mathbf{S}_u = [\mathbf{S}_u^1; ...; \mathbf{S}_u^M]$ is the concatenated features of all segments. This part has only the trend of intensity. Intensity in $\mathbf{S}_u^m$ increases or keeps the same. The other part is a small set of quantitatively labeled frames, $\mathcal{D}_f = \{\mathbf{x}_n, y_n\}_{n=1}^N$, where $\mathbf{x}_n \in \mathbb{R}^d$ is the raw feature vector of the n-th frame and $y_n \in \mathbb{R}$ is its intensity. $\mathbf{X}_l \in \mathbb{R}^{N \times d}$ is the concatenated features of all labeled frames and $\mathbf{Y}_l \in \mathbb{R}^N$ is their AU intensities. Frames in $\mathbf{X}_l$ are randomly selected frames rather than neighbor frames. Each row of $\mathbf{X}_l$ and $\mathbf{S}_u$ represents the features of a frame. Note that $\mathbf{X}_l$ is a subset of $\mathbf{S}_u$. For convenience, we denote them separately to avoid extra frame indexes. Let $\mathbf{B} \in \mathbb{R}^{K \times d}$ denote the basis vectors. K is the number of basis vectors. $\mathbf{\Phi}_l$ denotes the coefficients of labeled frames $\mathbf{X}_l$ in latent space, and $\mathbf{\Phi}_u = [\mathbf{\Phi}_u^1; ...; \mathbf{\Phi}_u^M]$ denotes the coefficients of segments $\mathbf{S}_u$, where $\mathbf{\Phi}_u^m \in \mathbb{R}^{T_m \times K}$. $\mathbf{w} \in \mathbb{R}^K$ denotes the parameters of the estimator. **For training**, given $\mathcal{D}_f$ and $\mathcal{D}_s$, we jointly learn the representation $\mathbf{\Phi}_l$ and $\mathbf{\Phi}_u$, the subspace span$(\mathbf{B})$, and the intensity estimator $\mathbf{w}$.

3.3. Hard constraints from human knowledge

Limited AU intensity annotations In training sequences, only few frames are labeled with AU intensities. The intensity labels provide strong supervision for joint learning. The representation and estimator are encouraged to satisfy

$$\mathbf{\Phi}_l \mathbf{w} = \mathbf{Y}_l. \tag{1}$$

The annotations of few frames are encoded as equality constraints. It is equivalent to put the loss in the objective, *i.e.*,

$$L_l(\mathbf{w}, \mathbf{\Phi}_l, \mathcal{D}_f) = \frac{\lambda_0}{2} \|\mathbf{\Phi}_l \mathbf{w} - \mathbf{Y}_l\|^2. \tag{2}$$

Temporal label ranking During a facial action, AU intensity evolves smoothly over time. As mentioned in Section 3.2, training sequences are split into segments according to key frames. AU intensity monotonically increases or keeps the same in each training segment. Though AU intensities of frames in a segment are unknown, the temporal relationships among multiple frames can provide weak supervision for joint learning. Instead of constraining the original representation $\mathbf{S}_u$, we emphasize domain knowledge on the new representation $\mathbf{\Phi}_u$. In a training segment, the representation and estimator are encouraged to satisfy that the intensity of the current frame should be larger than or equal to its previous frames,

$$\mathbf{\Phi}_u^{m,1}\mathbf{w} \leq ... \leq \mathbf{\Phi}_u^{m,i}\mathbf{w} \leq \mathbf{\Phi}_u^{m,i+1}\mathbf{w} \leq ... \leq \mathbf{\Phi}_u^{m,T_m}\mathbf{w},$$

where label ranking is encoded as inequality constraints. It is equivalent to

$$\mathbf{\Gamma}^m \mathbf{\Phi}_u^m \mathbf{w} \leq \mathbf{0}, \tag{3}$$

where $\mathbf{\Gamma}^m \in \mathbb{R}^{(T_m - 1) \times T_m}$ is a matrix with $\mathbf{\Gamma}_{i,i}^m = 1$, $\mathbf{\Gamma}_{i,i+1}^m = -1$, and other elements being 0's. For all qualitatively labeled segments, we have a set of constrains,

$$\mathbf{\Gamma}\mathbf{\Phi}_u \mathbf{w} \leq \mathbf{0}, \tag{4}$$

where $\mathbf{\Gamma} = \text{diag}([\mathbf{\Gamma}^1, \mathbf{\Gamma}^2, ..., \mathbf{\Gamma}^M])$ and $\mathbf{0}$ is a vector with all elements being 0's. Different from [43, 42], we emphasize ranking constrains on both representation and estimator.

Positive intensity AU intensity is a non-negative value. According to this prior knowledge, the prediction of AU intensity is encouraged to be not less than 0. Such knowledge is encoded as constraints which provide weak supervision for joint learning. The constraints are defined as

$$\mathbf{\Phi}_u \mathbf{w} \geq \mathbf{0}, \mathbf{\Phi}_l \mathbf{w} \geq \mathbf{0}. \tag{5}$$

The positive intensity is encoded as inequality constraints.

3.4. Soft constraints from human knowledge

Temporal label smoothness AU intensity is labeled according to corresponding local appearance. Since muscles move smoothly, facial appearance also changes smoothly over time. In a training segment, the intensity of a frame is close to the intensities of its neighbor frames. The representation and estimator are encouraged to satisfy that the intensities of neighbor frames should be similar, *i.e.*, $\|\mathbf{\Phi}_u^{m,i}\mathbf{w} - \mathbf{\Phi}_u^{m,j}\mathbf{w}\|^2$ is supposed to be small for two neighbor frames i and j. Considering all qualitatively labeled segments, we have the following regularization, *i.e.*,

$$\begin{aligned} R_I(\mathbf{w}, \mathbf{\Phi}_u, \mathcal{D}_s) &= \frac{1}{2}\sum_{m=1}^{M}\sum_{i,j}^{T_m}\mathbf{C}_{i,j}^m(\mathbf{\Phi}_u^{m,i}\mathbf{w} - \mathbf{\Phi}_u^{m,j}\mathbf{w})^2 \\ &= (\mathbf{\Phi}_u\mathbf{w})^T\mathbf{L}(\mathbf{\Phi}_u\mathbf{w}), \end{aligned} \tag{6}$$

where $\mathbf{L}^m = \mathbf{D}^m - \mathbf{C}^m$ and $\mathbf{L} = \text{diag}([\mathbf{L}^1, \mathbf{L}^2, ..., \mathbf{L}^M])$. $\mathbf{L}$ is a positive semi-definite matrix. $\mathbf{C}^m$ is an adjacent matrix, where $\mathbf{C}_{i,j}^m = 1$ if the j-th and i-th frames are neighbors. Otherwise $\mathbf{C}_{i,j}^m = 0$. $\mathbf{D}^m$ is a diagonal matrix with $\mathbf{D}_{i,i}^m = \sum_j \mathbf{C}_{i,j}^m$. Since frames in $\mathbf{X}_l$ are not neighbor frames, label smoothness can not be applied to $\mathbf{\Phi}_l\mathbf{w}$.

Temporal feature smoothness As facial muscles move smoothly, neighbor frames in sequences have similar facial appearance. The learned representation should keep such property that neighbor frames should have similar representations. The distance between representations of neighbor frames should small. Such knowledge can be encoded as a regularization term, *i.e.*,

$$\begin{aligned} R_F(\mathbf{\Phi}_u, \mathcal{D}_s) &= \frac{1}{2}\sum_{m=1}^{M}\sum_{i,j}^{T_m}\mathbf{C}_{i,j}^m||\mathbf{\Phi}_u^{m,i} - \mathbf{\Phi}_u^{m,j}||^2 \\ &= \text{tr}(\mathbf{\Phi}_u^T\mathbf{L}\mathbf{\Phi}_u), \end{aligned} \tag{7}$$

where $\mathbf{L}$ is the sames as Eq.(6). tr$(\cdot)$ represents the trace. Representation and estimator are coupled in Eqs. (6) and (4) while Eq. (7) involves only the representation.

3.5. Formulation

Given qualitatively labeled segments and limited quantitatively labeled frames, we formulate the problem as follows. For representation learning, learned coefficient matrix and basis vectors should be able to reconstruct raw features [23]. The reconstruction loss is defined as

$$\begin{aligned} &L_u(\mathbf{\Phi}_l, \mathbf{\Phi}_u, \mathbf{B}, \mathcal{D}_f, \mathcal{D}_s) \\ &= \frac{1}{2}\Big\| \begin{bmatrix}\mathbf{X}_l \\ \mathbf{S}_u\end{bmatrix} - \begin{bmatrix}\mathbf{\Phi}_l \\ \mathbf{\Phi}_u\end{bmatrix}\mathbf{B}\Big\|_F^2 + \lambda_1\Big\| \begin{bmatrix}\mathbf{\Phi}_l \\ \mathbf{\Phi}_u\end{bmatrix}^T\Big\|_{2,1}, \end{aligned} \tag{8}$$

where $\|\cdot\|_{2,1}$ encourages to learn features through the whole dataset rather than regularizing features of individual samples. To avoid degeneracy, the convex set for $\mathbf{B}$ is $\mathcal{B} = \{\mathbf{b} : \|\mathbf{b}\|_2 \leq 1\}$.

Considering unlabeled samples and human knowledge, the joint learning of image representation and intensity estimator can be formulated as

$$\begin{aligned} \min_{\mathbf{B}\in\mathcal{B}}\min_{\mathbf{w}}\min_{\mathbf{\Phi}_l,\mathbf{\Phi}_u} \quad & L_u(\mathbf{\Phi}_l, \mathbf{\Phi}_u, \mathbf{B}, \mathcal{D}_f, \mathcal{D}_s) + L_l(\mathbf{w}, \mathbf{\Phi}_l, \mathcal{D}_f) \\ & + \lambda_2 R_I(\mathbf{w}, \mathbf{\Phi}_u, \mathcal{D}_s) + \lambda_3 R_F(\mathbf{\Phi}_u, \mathcal{D}_s) \\ s.t. \quad & \mathbf{\Gamma}\mathbf{\Phi}_u\mathbf{w} \leq \mathbf{0}, \mathbf{\Phi}_l\mathbf{w} \geq \mathbf{0}, \mathbf{\Phi}_u\mathbf{w} \geq \mathbf{0}, \end{aligned} \tag{9}$$

where λ_0, λ_1, λ_2 and λ_3 are hyperparameters. The first term is the reconstruction error of all samples. The second is the loss of labeled samples. The third is the regularization of temporal label smoothness. The fourth is the regularization of temporal feature smoothness. The constraints represent the temporal label ranking and positive intensity.

Intensities of few frames provide strong supervision while domain knowledge provides weak supervision. Among types of knowledge, intensity and feature smoothness encourage smooth predictions. Label ranking encourages predictions in training segments to satisfy ordinal constraints, and positive intensity ensures the nonnegative prediction. We jointly learn the subspace $\mathbf{\Phi}$ and regressor $\mathbf{w}$, which are coupled through the knowledge. The soft and hard constraints involve both $\mathbf{\Phi}$ and $\mathbf{w}$. During optimization, the constraints and regularizations cooperate with each other to find the optimal solution of $\mathbf{\Phi}$ and $\mathbf{w}$.

3.6. Alternating optimization

Problem (9) is not jointly convex in all variables, but it is convex in each of them. Since Eq.(7) contains only $\mathbf{\Phi}_u$, we can not just optimize $\mathbf{\Phi}\mathbf{B}$ and $\mathbf{\Phi}\mathbf{w}$ by treating them as new variables. We propose an algorithm to solve the problem based on ADMM [3]. The scaled form of augmented Lagrangian function is

$$\begin{aligned} &L_{\boldsymbol{\rho}}(\mathbf{\Phi}, \mathbf{B}, \mathbf{w}, \mathbf{C}_., \mathbf{\Lambda}_., \mathbf{Z}_., \mathbf{V}_.) \\ = \; & \frac{1}{2}\Big\| \begin{bmatrix}\mathbf{X}_l \\ \mathbf{S}_u\end{bmatrix} - \begin{bmatrix}\mathbf{\Phi}_l \\ \mathbf{\Phi}_u\end{bmatrix}\mathbf{B}\Big\|_F^2 + \lambda_1\Big\| \begin{bmatrix}\mathbf{C}_l \\ \mathbf{C}_u\end{bmatrix}^T\Big\|_{2,1} \\ & + \frac{\rho_1}{2}\Big\| \begin{bmatrix}\mathbf{\Phi}_l \\ \mathbf{\Phi}_u\end{bmatrix} - \begin{bmatrix}\mathbf{C}_l \\ \mathbf{C}_u\end{bmatrix} + \begin{bmatrix}\mathbf{\Lambda}_l \\ \mathbf{\Lambda}_u\end{bmatrix}\Big\|_F^2 - \frac{\rho_1}{2}\Big\| \begin{bmatrix}\mathbf{\Lambda}_l \\ \mathbf{\Lambda}_u\end{bmatrix}\Big\|_F^2 \\ & + \mathbf{I}_-(\mathbf{Z}_0) + \frac{\rho_2}{2}||\mathbf{\Gamma}\mathbf{\Phi}_u\mathbf{w} - \mathbf{Z}_0 + \mathbf{V}_0||^2 - \frac{\rho_2}{2}||\mathbf{V}_0||^2 \\ & + \mathbf{I}_+(\mathbf{Z}_1) + \frac{\rho_3}{2}||\mathbf{\Phi}_l\mathbf{w} - \mathbf{Z}_1 + \mathbf{V}_1||^2 - \frac{\rho_3}{2}||\mathbf{V}_1||^2 \\ & + \mathbf{I}_+(\mathbf{Z}_2) + \frac{\rho_3}{2}||\mathbf{\Phi}_u\mathbf{w} - \mathbf{Z}_2 + \mathbf{V}_2||^2 - \frac{\rho_3}{2}||\mathbf{V}_2||^2 \\ & + \frac{\lambda_0}{2}||\mathbf{\Phi}_l\mathbf{w} - \mathbf{Y}_l||^2 + \lambda_2\mathbf{w}^T\mathbf{\Phi}_u\mathbf{L}^T\mathbf{\Phi}_u\mathbf{w} + \lambda_3\text{tr}(\mathbf{\Phi}_u\mathbf{L}^T\mathbf{\Phi}_u), \end{aligned} \tag{10}$$

where $\mathbf{\Phi}_. = \{\mathbf{\Phi}_l, \mathbf{\Phi}_u\}$, $\mathbf{C}_. = \{\mathbf{C}_l, \mathbf{C}_u\}$, $\mathbf{Z}_. = \{\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3\}$, and $\mathbf{V}_. = \{\mathbf{V}_0, \mathbf{V}_1, \mathbf{V}_2\}$. $\mathbf{C}_.$ and $\mathbf{Z}_.$ are introduced variables while $\mathbf{\Lambda}_.$ and $\mathbf{V}_.$ are the multipliers. They are introduced to handle the L21 norm and inequality constraints. $\boldsymbol{\rho} = \{\rho_1, \rho_2, \rho_3\}$ are penalty parameters to emphasize the importance of different knowledge. $\mathbf{I}_-(\cdot)$ and $\mathbf{I}_-(\cdot)$ are projection functions, *i.e.*, $\mathbf{I}_-(\cdot) = \min(\cdot, \mathbf{0})$ and $\mathbf{I}_+(\cdot) = \max(\cdot, \mathbf{0})$. We optimize each variable alternatively as follows (Algo. 1). PCA [37] is used to initialize $\mathbf{B}$, $\boldsymbol{\Phi}_l$, and $\boldsymbol{\Phi}_u$. $\mathbf{C}_l = \boldsymbol{\Phi}_l$ and $\mathbf{C}_u = \boldsymbol{\Phi}_u$ while other variables are randomly initialized. Note that currently updated variable will be used to update other variables. Following the conventional procedures of ADMM [3], the updates of the above variables are as follows:

$$\mathbf{B}^{(k+1)} \leftarrow \arg\min_{\mathbf{B}} L_{\boldsymbol{\rho}}(\cdots), \tag{11}$$

$$\boldsymbol{\Phi}_l^{(k+1)} \leftarrow \arg\min_{\boldsymbol{\Phi}_l} L_{\boldsymbol{\rho}}(\cdots), \tag{12}$$

$$\boldsymbol{\Phi}_u^{(k+1)} \leftarrow \arg\min_{\boldsymbol{\Phi}_u} L_{\boldsymbol{\rho}}(\cdots), \tag{13}$$

$$\mathbf{w}^{(k+1)} \leftarrow \arg\min_{\mathbf{w}} L_{\boldsymbol{\rho}}(\cdots), \tag{14}$$

$$\mathbf{C}_{\cdot}^{(k+1)} \leftarrow \arg\min_{\mathbf{C}_{\cdot}} L_{\boldsymbol{\rho}}(\cdots), \tag{15}$$

$$\mathbf{Z}_{\cdot}^{(k+1)} \leftarrow \arg\min_{\mathbf{Z}_{\cdot}} L_{\boldsymbol{\rho}}(\cdots), \tag{16}$$

$$\boldsymbol{\Lambda}_{\cdot}^{(k+1)} \leftarrow \boldsymbol{\Lambda}_{\cdot}^{(k)} + \boldsymbol{\Phi}_{\cdot}^{(k+1)} - \mathbf{C}_{\cdot}^{(k+1)}, \tag{17}$$

$$\mathbf{V}_0^{(k+1)} \leftarrow \mathbf{V}_0^{(k)} + \boldsymbol{\Gamma}\boldsymbol{\Phi}_u^{(k+1)}\mathbf{w}^{(k+1)} - \mathbf{Z}_0^{(k+1)}, \tag{18}$$

$$\mathbf{V}_1^{(k+1)} \leftarrow \mathbf{V}_1^{(k)} + \boldsymbol{\Phi}_l^{(k+1)}\mathbf{w}^{(k+1)} - \mathbf{Z}_1^{(k+1)}, \tag{19}$$

$$\mathbf{V}_2^{(k+1)} \leftarrow \mathbf{V}_2^{(k)} + \boldsymbol{\Phi}_u^{(k+1)}\mathbf{w}^{(k+1)} - \mathbf{Z}_2^{(k+1)}. \tag{20}$$

For **Problem** (11), we firstly obtain the closed-form solution of $\tilde{\mathbf{B}}$ by taking the gradient and setting it to 0, *i.e.*,

$$\tilde{\mathbf{B}} = [\boldsymbol{\Phi}_l^T\boldsymbol{\Phi}_l + \boldsymbol{\Phi}_u^T\boldsymbol{\Phi}_u]^{-1}[\boldsymbol{\Phi}_l^T\mathbf{X}_l + \boldsymbol{\Phi}_u^T\mathbf{X}_u]. \tag{21}$$

We project $\tilde{\mathbf{B}}$ into $\mathcal{B} = \{\mathbf{b} : ||\mathbf{b}||_2 = 1\}$ by normalizing each row of $\tilde{\mathbf{B}}$, *i.e.*, $\mathbf{B}_{i\cdot} = \frac{\tilde{\mathbf{B}}_{i\cdot}}{||\tilde{\mathbf{B}}_{i\cdot}||_2}$, where $\tilde{\mathbf{B}}_{i\cdot}$ is the i-th row of $\tilde{\mathbf{B}}$.

For **Problem** (12) and (14), we can get the closed-form solutions for $\boldsymbol{\Phi}_l$ and $\mathbf{w}$ by computing the gradient and setting the gradient to 0.

For **Problem** (13), though we can get the closed-form solution by taking the gradient of $\boldsymbol{\Phi}_u$, the computation is inefficient since it involves the inverse of a large matrix. Instead we use a gradient-based method to update $\boldsymbol{\Phi}_u$, *i.e.*,

$$\boldsymbol{\Phi}_u \leftarrow \boldsymbol{\Phi}_u - \alpha\nabla_u, \tag{22}$$

where ∇_u is the gradient of $\boldsymbol{\Phi}_u$ and the step size α is obtained by exact line search. Detailed representation of α is in the supplementary material.

For **Problem** (15), the subproblem with respect to $\mathbf{C}_l$ and $\mathbf{C}_u$ is

$$\min_{\mathbf{C}_l, \mathbf{C}_u} \lambda_1 \left\| \begin{bmatrix} \mathbf{C}_l \\ \mathbf{C}_u \end{bmatrix}^T \right\|_{2,1} + \frac{\rho_1}{2} \left\| \begin{bmatrix} \boldsymbol{\Phi}_l \\ \boldsymbol{\Phi}_u \end{bmatrix} - \begin{bmatrix} \mathbf{C}_l \\ \mathbf{C}_u \end{bmatrix} + \begin{bmatrix} \boldsymbol{\Lambda}_l \\ \boldsymbol{\Lambda}_u \end{bmatrix} \right\|_F^2.$$

Algorithm 1 Joint Representation and Estimator Learning.

```
Input: Labeled frames $\mathcal{D}_f$ and weakly labeled sequences $\mathcal{D}_s$. Penalty parameters $\{\lambda_i\}_{i=0}^3$ and $\{\rho_i\}_{i=1}^3$.
Output: Representation $\boldsymbol{\Phi}_l$ and $\boldsymbol{\Phi}_u$, basis vectors $\mathbf{B}$, and the estimator $\mathbf{w}$.
Initi: use PCA to obtain $\mathbf{B}$, $\boldsymbol{\Phi}_l$, and $\boldsymbol{\Phi}_u$. $\mathbf{C}_l = \boldsymbol{\Phi}_l$ and $\mathbf{C}_u = \boldsymbol{\Phi}_u$. Randomly initialize $\mathbf{Z}_\cdot$, $\boldsymbol{\Lambda}_\cdot$, and $\mathbf{V}_\cdot$.
while not converging do
    Update variables by solving Problem (11) $\sim$ (16)
    Update Lagrangian multipliers by Eqs. (17) $\sim$ (20)
end while
return $\boldsymbol{\Phi}_l$ , $\boldsymbol{\Phi}_u$, $\mathbf{B}$, and $\mathbf{w}$.
```

Let $\mathbf{C} = [\mathbf{C}_l; \mathbf{C}_u]$, $\boldsymbol{\Phi} = [\boldsymbol{\Phi}_l; \boldsymbol{\Phi}_u]$, and $\boldsymbol{\Lambda} = [\boldsymbol{\Lambda}_l; \boldsymbol{\Lambda}_u]$. The problem can be decomposed into small problems, *i.e.*,

$$\mathbf{C}_{\cdot i} = \arg\min_{\mathbf{C}_{\cdot i}} \lambda_1||\mathbf{C}_{\cdot i}||_2 + \frac{\rho_1}{2}||\boldsymbol{\Phi}_{\cdot i} - \mathbf{C}_{\cdot i} + \boldsymbol{\Lambda}_{\cdot i}||_F^2,$$

where $\mathbf{C}_{\cdot i}$ is the i-th column of $\mathbf{C}$, $\boldsymbol{\Phi}_{\cdot i}$ is the i-th column of $\boldsymbol{\Phi}$, and $\boldsymbol{\Lambda}_{\cdot i}$ is the i-th column of $\boldsymbol{\Lambda}$. The solution is

$$\mathbf{C}_{\cdot i} = S_{\lambda_1/\rho_1}(\boldsymbol{\Phi}_{\cdot i} + \boldsymbol{\Lambda}_{\cdot i}), \tag{23}$$

where $S_k(\mathbf{a}) = [1 - \frac{k}{||\mathbf{a}||_2}]_+ \odot \mathbf{a}$ and $S_k(0) = 0$. $[\cdot]_+ = \max(\cdot, 0)$. $\odot$ represents pairwise product.

For **Problem** (16), the solutions for $\mathbf{Z}_0$, $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are

$$\mathbf{Z}_0 = \min\{0, \boldsymbol{\Gamma}\boldsymbol{\Phi}_u\mathbf{w} + \mathbf{V}_0\}, \tag{24}$$

$$\mathbf{Z}_1 = \max\{0, \boldsymbol{\Phi}_l\mathbf{w} + \mathbf{V}_1\}, \tag{25}$$

$$\mathbf{Z}_2 = \max\{0, \boldsymbol{\Phi}_u\mathbf{w} + \mathbf{V}_2\}. \tag{26}$$

The optimization details of each subproblem are presented in the supplementary material.

For testing. We estimate intensities of testing samples in a transductive manner. Let $\mathbf{X}_t$ denote testing samples and $\boldsymbol{\Phi}_t$ denote their coefficients. We jointly learn the model and estimate intensities of testing samples by simply augmenting $\mathbf{S}_u$ and $\boldsymbol{\Phi}_u$, *i.e.*, $\mathbf{S} = [\mathbf{S}_u; \mathbf{X}_t]$ and $\boldsymbol{\Phi} = [\boldsymbol{\Phi}_u; \boldsymbol{\Phi}_t]$. $\mathbf{L} = \text{diag}(\mathbf{L}^1, ..., \mathbf{L}^M, \mathbf{L}^t)$ and $\boldsymbol{\Gamma} = \text{diag}(\boldsymbol{\Gamma}^1, ..., \boldsymbol{\Gamma}^M, \boldsymbol{\Gamma}^t)$, where $\mathbf{L}^t$ and $\boldsymbol{\Gamma}^t$ are matrices with all elements being 0's because we have no information about testing samples and knowledge is only applied to training segments. During testing, we perform frame-level prediction by using $\mathbf{Y}_t = \boldsymbol{\Phi}_t\mathbf{w}$.

4. Experiments

4.1. Settings

Data. BP4D-spontaneous database [39] was used as the Train/Development splits of the FERA 2015 Challenge [32]. AU intensity is qualified into 6 discrete levels. Following the protocol of FERA 2015, we use the Train split for training and the Development split for evaluation.

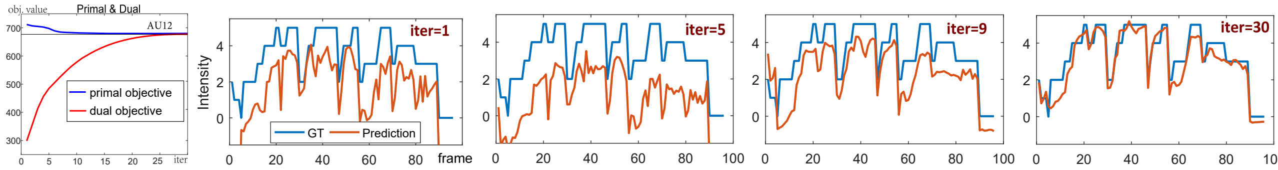

Figure 2: The learning curve of KJRE on AU12 under the scenario that 6% of training samples are annotated. The last four show the intensity prediction on a testing sequence at different iterations.

Note that FERA 2017 [33] is a challenge for AU intensity estimation under different poses. Since our goal is to learn an estimator with limited annotations, we use FERA 2015. DISFA [19] consists of 27 sequences from 27 subjects. We perform 5-fold subject independent cross validation. For feature extraction, we follow the same procedures in [32, 33, 42] to extract 218D features.

Annotation. Instead of using the intensity label of each frame, our method needs only weak annotation (see Sec. 3.1), *i.e.,* identifying locations of key frames and labeling the intensities of few frames. We follow [17] to identify key frames. Knowledge can be applied to all training segments even though no frame has intensity annotation. Since sequences in both databases are captured with a high frame rate, faces in consecutive frames have minor changes. Sequences are downsampled by selecting one frame every five frames. Segment length varies between 10 and 80. Distributions of AU intensity are shown in Fig. 3 . For evaluation, we vary the proportion of labeled frames in the training set, including 2%, 4%, 6%, 8%, 10%, 30%, 50%, 70%, 90%, and 100%. Labeled frames are randomly selected and key frames have the priority to be selected. We perform each experiment 5 times and report the average performance.

Evaluation metrics. Pearson Correlation Coefficient (PCC), Intra-Class Correlation (ICC(3,1) [28]), and Mean Absolute Error (MAE) are adopted as the measures for evaluation. K, $\{\lambda_i\}_{i=0}^3$ and $\{\rho_i\}_{i=1}^3$ are the hyperparameters of our model. For parameter selection, the training set is divided into two parts with 60% segments for training and 40% for validation. We use grid search strategy to find the best hyperparameters from $K \in \{60, 80, 100, 120, 140\}$, $\{\lambda_i\}_{i=0}^3, \{\rho_i\}_{i=1}^3 \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$.

Models. We incorporate four types of human knowledge to jointly learn representation and estimator (KJRE). To verify the effectiveness of each type of knowledge, we compare the performance of not using knowledge (JRE) with the performance of using only one type of knowledge, including label ranking (KJRE-O), label smoothness (KJRE-I), feature smoothness (KJRE-F), and positive intensity (KJRE-P). KE-PCA first uses PCA to get the representation and then uses knowledge for estimator learning. We then compare with the state-of-the-art supervised methods (SVR [32], RVR [10], SOVRIM [5], LT [11], COR [35], DSRVM [12]), semi-supervised methods (CSL [40], HSSR [13]), and weakly supervised methods (OSVR [43], BORMIR [42]). Supervised methods use only labeled samples while weakly and semi-supervised methods use both labeled and unlabeled samples. For weakly supervised methods, OSVR, BORMIR, and our method require preprocessing by splitting sequences into segments. We also compare to supervised deep models (CCNN [34], 2CD [31]) and a weakly supervised deep model (KBSS [41]).

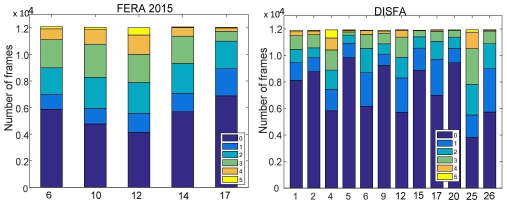

Figure 3: Intensity distribution.

Complexity and convergence. The computational complexity is $\mathcal{O}(d^3 + T(K^3 + (2N + d)K^2 + NKd))$. The space complexity is $\mathcal{O}(Nd + d^2)$. d denotes the dimension of input space. T denotes the iterations of ADMM. When d and K are large, the complexity can be reduced to $\mathcal{O}(d^{2.373} + T(K^{2.373} + (2N + d)K^2 + NKd))$ by using [6] to compute matrix inversion. Fig. 2 illustrates the learning curve of AU12 with KJRE and also the prediction on a testing sequence at different iterations. The primal objective decreases while the dual objective increases. When they get close, the algorithm converges. Our method has the same complexity as BORMIR [42] ($O(Nd^2T)$) and is more efficient than OSVR [43] ($O(N^2dT)$) when $N \gg d$.

4.2. Results

Comparison with baseline methods. The results are shown in Table 1. Methods are valuated under the scenario that 6% of training frames have intensity labels. Each method achieves better performance on FERA than on DISFA because DISFA is a more challenge database due to the low-quality images, large head poses, complex illuminations, and imbalanced intensity distribution. Detailed analyses are as follows. Firstly, methods that use one type of human knowledge, including KJRE-O, KJRE-I, KJRE-F, and KJRE-P, achieve better results than JRE which does not use any type of knowledge. It demonstrates the effectiveness of each type of knowledge. Label ranking and label smoothness are relatively more important than feature smoothness and positive intensity. Secondly, KJRE combines all types of knowledge and achieves better performance than JRE as well as methods that use partial knowledge. It further

Table 1: Comparison with baseline methods. The performance is evaluated under the scenario that 6% of training frames are annotated. Bold numbers with gray background indicate the best performance; bold numbers indicate the second best.

Database		FERA 2015						DISFA												
AU		6	10	12	14	17	Avg	1	2	4	5	6	9	12	15	17	20	25	26	Avg
PCC	KE-PCA	.63	.49	.64	.18	**.45**	.48	.18	.29	.15	.34	.25	.32	.54	.10	.15	**.30**	.47	.27	.28
	JRE	.66	.56	.81	.30	.36	.54	.07	.26	.25	.29	.34	.29	.44	.05	.17	.17	.58	.22	.26
	KJRE-O	**.70**	**.60**	**.85**	**.39**	.37	**.58**	.14	.30	**.31**	**.39**	.43	**.35**	**.62**	**.12**	**.22**	.19	.70	**.28**	**.34**
	KJRE-I	.68	.60	.83	.36	.36	.57	.15	.30	.27	**.36**	.43	**.34**	.60	.10	**.20**	.21	**.71**	**.29**	.33
	KJRE-F	.67	.57	.82	.32	.37	.55	.11	.30	.26	.35	.39	.31	.49	.07	.19	.19	.64	.25	.30
	KJRE-P	.69	.60	.82	.33	.35	.56	**.21**	**.38**	**.28**	.28	**.52**	.33	.59	.04	.10	.20	.69	.25	.32
	KJRE	**.72**	**.65**	**.87**	**.40**	**.43**	**.62**	**.28**	**.38**	.26	.34	**.54**	.33	**.69**	**.18**	.18	**.22**	**.75**	.25	**.37**
ICC	KE-PCA	.61	.45	.63	.15	**.39**	.45	.05	.07	.03	.13	.06	.09	.22	.03	.04	.07	.17	.06	.08
	JRE	.65	.56	.81	.30	.36	.54	.07	.22	.17	.22	.31	.28	.33	.06	.13	.16	.52	.19	.22
	KJRE-O	**.70**	**.60**	**.85**	**.37**	.36	**.58**	.14	.27	**.26**	**.37**	.42	**.34**	**.61**	**.11**	**.21**	**.18**	.70	**.27**	**.32**
	KJRE-I	.68	.59	.83	.35	.36	.56	.15	.29	**.27**	**.35**	.41	**.31**	.60	.09	**.20**	.18	**.70**	**.28**	.32
	KJRE-F	.66	.57	.81	.32	.37	.55	.11	.28	.23	.33	.37	.29	.41	.08	.18	.18	.61	.23	.27
	KJRE-P	.69	.58	.82	.32	.34	.55	**.20**	**.33**	.26	.26	**.48**	.30	.57	.04	.11	.18	.68	.24	.30
	KJRE	**.71**	**.61**	**.87**	**.39**	**.42**	**.60**	**.27**	**.35**	.25	.33	**.51**	.31	**.67**	**.14**	.17	**.20**	**.74**	.25	**.35**
MAE	KE-PCA	1.56	2.02	2.37	1.78	1.09	1.76	**.81**	**.64**	**1.51**	**.41**	1.03	**.61**	1.21	**.43**	**.77**	**.44**	1.92	**.94**	**.89**
	JRE	1.07	1.09	.87	1.23	1.04	1.06	1.75	1.59	2.97	1.15	1.38	1.39	2.05	1.00	1.33	1.00	1.49	1.57	1.56
	KJRE-O	**.91**	**1.00**	**.71**	**1.10**	**.92**	**.93**	1.38	1.31	2.28	.74	.96	1.09	1.01	.71	.92	.90	.97	1.28	1.13
	KJRE-I	.98	1.00	.78	1.14	.97	.97	1.08	.96	1.90	**.63**	.90	.97	**.95**	.66	.86	**.65**	**.91**	1.03	.96
	KJRE-F	1.09	1.06	.85	1.21	1.04	1.05	1.43	1.21	2.36	.81	1.16	1.25	1.68	.85	1.03	.79	1.22	1.38	1.26
	KJRE-P	.99	1.01	.78	1.15	.94	.98	1.21	.99	1.86	.90	**.89**	1.10	1.13	.71	1.06	.83	1.07	1.11	1.07
	KJRE	**.82**	**.95**	**.64**	**1.08**	**.85**	**.87**	**1.02**	**.92**	**1.86**	.70	**.79**	**.87**	**.77**	**.60**	**.80**	.72	**.96**	**.94**	**.91**

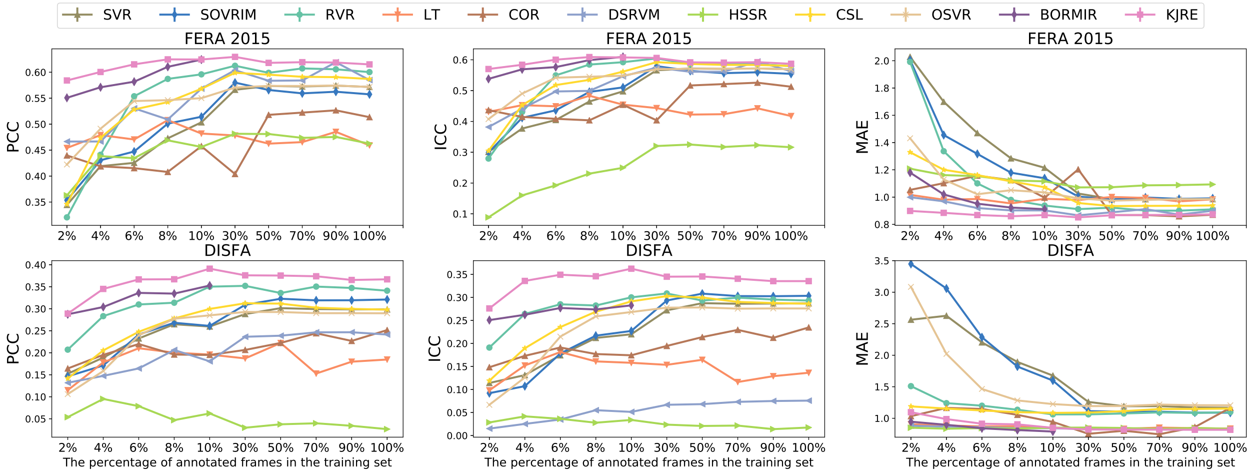

Figure 4: Comparison with the state-of-the-art methods under different annotation ratios. It presents the average performance under different annotation ratios. 'The percentage of annotated frames' refers to $\frac{\text{\#quantitatively annotated frames}}{\text{\#all frames}}$. Competitive methods are SVR [32], RVR [10], SOVRIM [5], LT [11], COR [35], DSRVM [12], CSL [40], HSSR [13], OSVR [43], and BORMIR [42].

demonstrates human knowledge helps improve both representation and estimator learning. Thirdly, our method achieves much better results than KE-PCA which uses PCA to learn representation in an unsupervised manner and then incorporates knowledge for estimator learning. On DISFA, KE-PCA achieves slightly better MAE than our method, but it gets much worse PCC and ICC. Since intensity levels are imbalanced and the majority intensity is 0 in DISFA, the representation learning will be dominated by samples with the intensity of 0. It makes KE-PCA tend to predict the intensity of 0 for all testing samples. As the majority intensity is 0, KE-PCA can get good MAE, but poor performance in ICC. The comparison to KE-PCA shows that our method is more robust on the learning of representation and estimator when the database is imbalanced.

Comparison with the state-of-the-art methods. Fig. 4 presents the average performance of methods under different annotation ratios. Table 2 shows the results when the annotation ratios is 6%. The state-of-the-art methods are evaluated by using the code provided by the authors. Note that BORMIR [42] can not use segments that have no intensity annotations of peak and valley frames. It can use at most about 10% labeled frames. As shown in Fig. 4 and Table 2, on FERA 2015, our method achieves better performance than other methods under all evaluation metrics, especially when the annotation ratio is small such as 2% ∼ 10%. On DISFA, our method achieves better performance than other methods under PCC and ICC. Its MAE

Table 2: Comparison with the state-of-the-art methods. The performance is evaluated under the scenario that 6% of training frames are annotated. Bold numbers with gray background indicate the best performance; bold numbers indicate the second best.

Database		FERA 2015						DISFA												
AU		6	10	12	14	17	Avg	1	2	4	5	6	9	12	15	17	20	25	26	Avg
PCC	SVR [32]	.45	.42	.74	.25	.28	.43	.06	.30	.21	.29	.33	.15	.44	.15	.01	.16	.54	.14	.23
	SOVRIM [5]	.50	.43	.76	.24	.30	.45	.06	.29	.24	.27	.36	.10	.41	**.23**	.06	.16	.57	.14	.24
	RVR [10]	.67	.60	.82	.27	.41	.55	**.25**	.35	.04	**.31**	.40	.27	**.61**	.11	**.22**	.06	**.82**	**.27**	.31
	LT [11]	.59	.61	.71	.35	.09	.47	.22	.07	.04	.24	.42	.13	.45	.04	.13	.00	.52	**.28**	.21
	COR [35]	.49	.54	.69	.16	.18	.42	.18	.18	**.29**	.04	.32	.24	.39	.09	.12	.01	.66	.13	.22
	DSRVM [12]	.63	**.64**	.81	**.36**	.22	.53	.12	.02	.12	.06	.37	.20	.40	.07	.07	-.04	.52	.06	.16
	HSSR [13]	.53	.55	.63	.26	.20	.43	.05	.01	-.02	.04	.11	.14	.16	.00	.00	.09	.25	.10	.08
	CSL [40]	.66	.52	.76	.28	**.42**	.53	.16	.27	.28	.23	.34	.22	.37	.18	.06	.15	.53	.19	.25
	OSVR [43]	.63	.54	**.84**	.35	.37	.55	.16	.23	.20	.12	.29	**.29**	.47	.14	.07	.14	.59	.17	.24
	BORMIR [42]	**.71**	.63	.82	.32	**.42**	**.58**	.24	**.36**	**0.37**	.21	**.47**	.24	.60	**.31**	**.20**	**.20**	.70	.13	**.34**
	KJRE	**.72**	**.65**	**.87**	**.40**	**.43**	**.62**	**.28**	**.38**	.26	**.34**	**.54**	**.33**	**.69**	.18	.18	**.22**	**.75**	.25	**.37**
ICC	SVR [32]	.40	.41	.73	.23	.26	.41	.04	.24	.16	.18	.27	.12	.32	.11	.00	.11	.47	.07	.17
	SOVRIM [5]	.47	.43	.76	.23	.29	.44	.05	.22	.19	.16	.30	.08	.28	**.16**	.04	.10	.49	.07	.18
	RVR [10]	.65	.59	.82	.27	.41	.55	**.24**	**.31**	.04	**.29**	.35	.23	**.59**	.08	**.18**	.05	**.81**	**.27**	**.29**
	LT [11]	.58	.58	.69	.32	.08	.45	.18	.05	.01	.19	.39	.09	.42	.02	.10	.00	.51	.21	.18
	COR [35]	.48	.54	.69	.16	.18	.41	.14	.12	.24	.01	.29	.19	.38	.05	.10	.01	.65	.11	.19
	DSRVM [12]	.60	**.62**	.80	.31	.16	.50	.02	.00	.01	.00	.07	.04	.08	.02	.02	-.02	.16	.00	.03
	HSSR [13]	.28	.24	.28	.09	.07	.19	.00	-.02	-.03	.03	.04	.06	.11	-.01	.01	.05	.14	.08	.04
	CSL [40]	.65	.49	.76	.27	**.42**	.52	.15	.26	**.27**	.22	.31	.21	.36	**.18**	.05	**.15**	.49	.18	.24
	OSVR [43]	.63	.53	**.84**	**.35**	.36	.54	.16	.23	.17	.11	.26	**.28**	.37	.14	.05	.14	.56	.12	.22
	BORMIR [42]	**.71**	**.62**	.82	.31	**.42**	**.58**	.19	.27	**.33**	.14	**.41**	.18	.58	**.16**	.16	.09	.69	.13	.28
	KJRE	**.71**	.61	**.87**	**.39**	**.42**	**.60**	**.27**	**.35**	.25	**.33**	**.51**	**.31**	**.67**	.14	**.17**	**.20**	**.74**	**.25**	**.35**
MAE	SVR [32]	1.71	1.44	1.06	1.76	1.38	1.47	2.42	1.94	3.41	2.07	1.83	2.28	2.31	1.54	2.45	1.76	1.95	2.47	2.20
	SOVRIM [5]	1.41	1.46	.97	1.54	1.21	1.32	2.43	1.91	3.52	2.31	1.90	2.22	2.60	1.60	2.43	1.80	1.86	2.85	2.29
	RVR [10]	1.16	1.13	.84	1.35	1.03	1.10	1.22	1.09	2.66	.88	1.03	1.20	1.31	.72	1.22	.79	**.78**	1.49	1.20
	LT [11]	.94	.99	.91	**1.09**	1.00	.99	1.00	.93	1.43	.61	.80	.84	**.91**	**.56**	**.73**	.67	1.13	.85	.87
	COR [35]	1.21	1.15	.88	1.63	.92	1.16	1.16	1.24	**1.18**	2.88	.98	.91	1.11	.91	**.68**	1.15	**.77**	**.80**	1.15
	DSRVM [12]	.89	**.96**	.78	**1.09**	**.87**	**.92**	**.89**	**.82**	1.31	.64	.88	**.78**	.98	**.56**	.77	**.58**	1.27	.82	.86
	HSSR [13]	1.16	1.24	1.25	1.21	.92	1.16	.92	.84	**1.30**	**.44**	.92	**.66**	1.08	**.53**	**.73**	**.59**	1.31	**.80**	**.84**
	CSL [40]	.97	1.37	1.22	1.32	.94	1.16	1.20	1.16	1.40	1.11	.98	1.19	1.03	.98	1.17	1.05	1.08	1.12	1.12
	OSVR [43]	.99	1.13	**.77**	1.15	1.06	1.02	1.44	1.14	3.09	.90	1.61	.94	1.94	.78	1.75	.85	1.36	1.78	1.46
	BORMIR [42]	**.83**	.99	.92	1.14	0.90	0.96	**.90**	**.77**	1.48	**.58**	**.78**	**.78**	**0.91**	**0.56**	.75	0.64	.95	.98	**.84**
	KJRE	**.82**	**.95**	**.64**	**1.08**	**.85**	**.87**	1.02	.92	1.86	.70	**.79**	.87	**.77**	.60	.80	.72	.96	.94	.91

Table 3: Comparison with deep models

Database	FERA 2015			DISFA		
Method	PCC	ICC	MAE	PCC	ICC	MAE
# annotated frames: more than 75,000						
CCNN [34]*	-	.63	1.26	-	.38	.66
2DC [31]*	-	.66	-	-	.50	-
# annotated frames: 120						
KBSS [41]	.50	.49	1.17	.23	.21	**.68**
KJRE	**.58**	**.57**	**.90**	**.29**	**.28**	1.10
# annotated frames: 360						
KBSS [41]	.60	.58	.94	.31	.30	**.62**
KJRE	**.62**	**.60**	**.87**	**.37**	**.35**	.91
# annotated frames: 1000						
KBSS [41]	**.63**	**.63**	.88	.37	.35	**.46**
KJRE	**.63**	.61	**.85**	**.38**	**.36**	.82

is slightly worse than HSSR, LT, DSRVM, and BORMIR when annotation ratio is less than 10%. HSSR, LT, and D-SRVM have good performance in MAE because they are sensitive to the imbalanced AU intensity distribution during learning and tend to predict intensity level 0 for all testing samples. Since the majority intensity is 0, this makes them have good MAE but poor ICC and MAE.

Table 3 shows the comparison with deep models when using limited annotated frames. The average performance is presented. Results of CCNN and 2DC are adapted from the corresponding papers. On FERA 2015, our method outperforms KBSS when using only 120 or 360 annotated frames. When using 1000 frames, our method is comparative to KB-SS. On DISFA, our method achieves better PCC and ICC. The results show that our method is applicable to databases with scarce annotated frames, even with few hundreds of annotated frames where our method can outperform KBSS.

5. Conclusion

We propose a general framework for AU intensity estimation which jointly learns representation and estimator with limited annotations. Besides, it can flexibly incorporate various types of human knowledge. Human knowledge is used to provided weak supervision for the joint learning and to efficiently exploit unlabeled images. We also propose an algorithm for optimization based on the framework of ADMM. Evaluations on two benchmark databases demonstrate the effectiveness of the proposed method, especially when the ratio of intensity annotations is small.

Acknowledgments: This work is partially supported by the National Key R&D Program of China (Grant No. 2018YFC0807500) and by NSFC Nos. 61832016 and 61720106006. Qiang Ji's involvement in this work is supported in part by the US National Science Foundation award CNS No. 1629856.

References

[1] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR*, 2006. 2

[2] B. Bentsianov and A. Blitzer. Facial anatomy. *Clinics in dermatology*, 2004.

[3] S. Boyd. Alternating direction method of multipliers. 4, 5

[4] O. Chapelle and A. Zien. Semi-supervised classification by low density separation. In *AISTATS*, 2005. 2

[5] W. Chu and S. S. Keerthi. New approaches to support vector ordinal regression. In *ICML*, 2005. 6, 7, 8

[6] D. Coppersmith and S. Winograd. Matrix multiplication via arithmetic progressions. In *Proceedings of the nineteenth annual ACM symposium on Theory of computing*, pages 1–6. ACM, 1987. 6

[7] P. Ekman and W. V. Friesen. *Manual for the facial action coding system*. Consulting Psychologists Press, 1978. 1

[8] A. Gudi, H. E. Tasli, T. M. den Uyl, and A. Maroulis. Deep learning based facs action unit occurrence and intensity estimation. In *FG workshop*, 2015. 1, 2

[9] R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. 2000. 2

[10] S. Kaltwang, O. Rudovic, and M. Pantic. Continuous pain intensity estimation from facial expressions. In *ISVC*, 2012. 1, 2, 6, 7, 8

[11] S. Kaltwang, S. Todorovic, and M. Pantic. Latent trees for estimating intensity of facial action units. In *CVPR*, 2015. 2, 6, 7, 8

[12] S. Kaltwang, S. Todorovic, and M. Pantic. Doubly sparse relevance vector machine for continuous facial behavior estimation. *TPAMI*, 2016. 2, 6, 7, 8

[13] K. I. Kim, F. Steinke, and M. Hein. Semi-supervised regression using hessian energy with an application to semi-supervised dimensionality reduction. In *NIPS*, 2009. 2, 6, 7, 8

[14] Y. Li, S. M. Mavadati, M. H. Mahoor, Y. Zhao, and Q. Ji. Measuring the intensity of spontaneous facial action units with dynamic bayesian network. *PR*, 2015. 2

[15] W. Liu, J. Wang, and S.-F. Chang. Robust and scalable graph-based semisupervised learning. *Proceedings of the IEEE*, 2012. 2

[16] 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 *CVPRW*, 2009. 1, 2

[17] M. Mavadati, P. Sanger, and M. H. Mahoor. Extended disfa dataset: Investigating posed and spontaneous facial expressions. In *CVPRW*, 2016. 2, 6

[18] S. M. Mavadati and M. H. Mahoor. Temporal facial expression modeling for automated action unit intensity measurement. In *ICPR*, 2014. 2

[19] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 2013. 6

[20] S. Melacci and M. Belkin. Laplacian support vector machines trained in the primal. *JMLR*, 2011. 2

[21] Z. Ming, A. Bugeau, J.-L. Rouas, and T. Shochi. Facial action units intensity estimation by the fusion of features with multi-kernel support vector machine. In *FG Workshop*, volume 6, pages 1–6. IEEE, 2015. 2

[22] J. Nicolle, K. Bailly, and M. Chetouani. Facial action unit intensity prediction via hard multi-task metric learning for kernel regression. In *FG Workshop*, volume 6, pages 1–6. IEEE, 2015. 2

[23] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997. 4

[24] O. Rudovic, V. Pavlovic, and M. Pantic. Context-sensitive dynamic ordinal regression for intensity estimation of facial action units. *TPAMI*, 2015. 2

[25] A. Ruiz, J. Van de Weijer, and X. Binefa. Regularized multi-concept mil for weakly-supervised facial behavior categorization. In *BMVC*, 2014. 2

[26] O. R. Ruiz, Adria, X. binefa, and M. Pantic. Multi-instance dynamic ordinal random fields for weakly-supervised facial behavior analysis. *arXiv preprint arXiv:1803.00907*. 1, 2

[27] G. Sandbach, S. Zafeiriou, and M. Pantic. Markov random field structures for facial action unit intensity estimation. In *ICCVW*, 2013. 2

[28] P. E. Shrout and J. L. Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 1979. 6

[29] K. Sikka, A. Dhall, and M. Bartlett. Weakly supervised pain localization using multiple instance learning. In *FG workshop*, 2013. 2

[30] K. Sikka, G. Sharma, and M. Bartlett. Lomo: Latent ordinal model for facial analysis in videos. In *CVPR*, 2016. 2

[31] D. L. Tran, R. Walecki, S. Eleftheriadis, B. Schuller, M. Pantic, et al. Deepcoder: Semi-parametric variational autoencoders for facial action unit intensity estimation. In *ICCV*, 2017. 2, 6, 8

[32] 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. In *FG workshop*, 2015. 2, 5, 6, 7, 8

[33] M. F. Valstar, E. Sánchez-Lozano, J. F. Cohn, L. A. Jeni, J. M. Girard, Z. Zhang, L. Yin, and M. Pantic. Fera 2017-addressing head pose in the third facial expression recognition and analysis challenge. In *FG*, 2017. 6

[34] R. Walecki, V. Pavlovic, B. Schuller, M. Pantic, et al. Deep structured learning for facial action unit intensity estimation. In *CVPR*, 2017. 2, 6, 8

[35] R. Walecki, O. Rudovic, M. Pantic, and V. Pavlovic. Copula ordinal regression for joint estimation of facial action unit intensity. In *CVPR*, 2016. 2, 6, 7, 8

[36] S. Wang, J. Yang, Z. Gao, and Q. Ji. Feature and label relation modeling for multiple-facial action unit classification and intensity estimation. *PR*, 2017. 2

[37] S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987. 5

[38] Z. Zafar and N. A. Khan. Pain intensity evaluation through facial action units. In *ICPR*, 2014. 2

[39] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *IVC*, 2014. 5

[40] X. Zhang, Y. Yu, M. White, R. Huang, and D. Schuurmans. Convex sparse coding, subspace learning, and semi-supervised extensions. In *AAAI*, 2011. 2, 3, 6, 7, 8

[41] Y. Zhang, W. Dong, B.-G. Hu, and Q. Ji. Weakly-supervised deep convolutional neural network learning for facial action unit intensity estimation. In *CVPR*, 2018. 1, 2, 6, 8

[42] Y. Zhang, R. Zhao, W. Dong, B.-G. Hu, and Q. Ji. Bilateral ordinal relevance multi-instance regression for facial action unit intensity estimation. In *CVPR*, 2018. 1, 2, 3, 6, 7, 8

[43] R. Zhao, Q. Gan, S. Wang, and Q. Ji. Facial expression intensity estimation using ordinal information. In *CVPR*, 2016. 1, 2, 3, 6, 7, 8