Facial Expression Intensity Estimation Using Ordinal Information

Rui Zhao[1], Quan Gan[2], Shangfei Wang[2]*, Qiang Ji[1]
[1]Department of Electrical, Computer & Systems Engineering, Rensselaer Polytechnic Institute
[2]School of Computer Science and Technology, University of Science and Technology of China
[1]{zhaor,jiq}@rpi.edu, [2]{gqquan@mail.,sfwang@}ustc.edu.cn

Abstract

Previous studies on facial expression analysis have been focused on recognizing basic expression categories. There is limited amount of work on the continuous expression intensity estimation, which is important for detecting and tracking emotion change. Part of the reason is the lack of labeled data with annotated expression intensity since expression intensity annotation requires expertise and is time consuming. In this work, we treat the expression intensity estimation as a regression problem. By taking advantage of the natural onset-apex-offset evolution pattern of facial expression, the proposed method can handle different amounts of annotations to perform frame-level expression intensity estimation. In fully supervised case, all the frames are provided with intensity annotations. In weakly supervised case, only the annotations of selected key frames are used. While in unsupervised case, expression intensity can be estimated without any annotations. An efficient optimization algorithm based on Alternating Direction Method of Multipliers (ADMM) is developed for solving the optimization problem associated with parameter learning. We demonstrate the effectiveness of proposed method by comparing it against both fully supervised and unsupervised approaches on benchmark facial expression datasets.

1. Introduction

Facial expression provides rich information in understanding a person's emotional state, feeling and attitude (see [8] for a survey). So far the majority of expression analysis work focus on recognition of basic expression categories including anger, happy, fear, surprise, sadness, disgust, contempt, *etc.* Recently, there is an increasing interest in a more fine-grained analysis, namely the facial expression intensity estimation. For instance, the pain intensity is used to evaluate the extent of discomfort in a healthcare application [13].

Automatic expression intensity estimation is challenging partially due to the lack of standard rule for expression intensity labeling. One way to define expression intensity uses the intensity of Action Units (AUs), which is a set of atomic facial muscle actions defined by Facial Action Coding System (FACS). [6]. For example, Prkachin and Solomon [21] used the intensity summation of four AUs: brow lowering, orbital tightening, levator contraction and eye closure to define the pain intensity. However, manually recognizing AU and annotating its intensity is a time consuming process with requirement of domain expertise. While automatic AU intensity estimation is still an open research problem [25, 15, 24].

Another way to define expression intensity uses relative difference between facial images presented at different stages of an expression. For example, Hess *et al.* [11] defined the expression intensity as the relative degree of displacement away from a neutral or relaxed facial expression. Dhall and Goecke [5] divided the dynamic process of smile into 6 stages. Despite the simple definition, there is no accurate way of determining different intensity levels except manual labeling, which requires substantial labor and expertise. Due to this reason, there are few datasets that come with expression intensity labels. One exception is [18].

In this work, we introduce a method that learns a frame-level expression intensity estimator by exploiting the ordinal information among different frames and intensity labels of selected frames, if available, in the training image sequences. Our approach is based on the observation that the temporal evolution of facial expression usually follows a particular order. Starting from a neutral stage where no expression is observed, we consider the expression intensity reaching its lowest level. Then we observe the onset of expression followed by apex, when the intensity reaches its peak. After reaching its peak, the expression intensity starts to reduce until it is back to neutral status. Even though the duration of each stage may vary under different occasions for different subjects, the general trend of the evolution remains the same. Figure 1 shows some expression sequence examples with such evolution trend. Similar idea of temporal evolution pattern has been applied to other computer

*Corresponding author

vision task such as action recognition [9]. Obtaining the labels for apex and onset/offset frames is usually less expensive. Some dataset has the setting of only recording the expression changing from onset to apex, which readily provides us with some relative intensity information.

Figure 1. Sample expression sequences from top: UNBC-McMaster shoulder pain [18], middle: CK+ [17] and bottom: BU-4DFE [32] datasets.

Existing work on intensity estimation usually fall at two ends of machine learning paradigms. In a fully supervised setting, Lee and Xu [14] adopted definition of the expression intensity in [11] and used Support Vector Regression (SVR) to model the facial expression intensity. However the model is subject dependent and needs to be trained for different subjects. More recently, Kaltwang *et al.* [13] performed continuous pain intensity estimation using Relevance Vector Regression (RVR) without considering ordinal information of the expression change. Rudovic *et al.* [22] proposed a Conditional Random Field (CRF) based model to combine the topology and ordinal state of facial affect data for joint expression recognition and intensity estimation. However, the inference is performed at the sequence level and the output only support discrete intensity values. Some follow-up works by the same authors [23, 24] learned pain intensity estimator in a fully supervised manner where the intensity labels of all the frames are provided.

On the unsupervised setting, Yang *et al.* [31] proposed a RankBoost based framework which learns a ranking model using ordinal relationship among image frames to do intensity estimation and expression recognition. However, a specially ordered sequence is required to be constructed for model learning and the relative intensity level among different sequences are different. Whitehill *et al.* [29] trained a binary classifier using GentleBoost for smile detection where the output score of classifier is used for intensity estimation. The learning is performed using individual images, where no ordinal information of an expression is used.

Our contributions include the following aspects. First, we propose a regression approach for expression intensity estimation which exploits both ordinal relationship among different frames within an expression sequence and absolute intensity labels if available. Second, we introduce a unified max-margin learning framework to simultaneously exploit the two sources of information. An efficient algorithm to solve the optimization problem is developed. Third, our method can generalize to different learning settings depending on the availability of expression intensity annotations.

As for the rest of this paper, a formal definition of the problem and used assumptions are described in section 2. We review some basic components of our method in section 3 and explain the proposed method in details in section 4. Experimental evaluation is provided in section 5, followed by conclusion and future work in section 6.

2. Problem Statement

In this section, we define the problem and state the assumptions we used. Our goal is to learn frame-level expression intensity estimator using expression sequences as training data with or without any intensity annotations.

Denote an expression sequence as $\mathbf{X} = \{\mathbf{x}_i \in \mathbb{R}^d | i = 1, ..., |\mathbf{X}|\}$, where $\mathbf{x}_i$ is the i^{th} individual frame, d is the feature dimension of a frame and $|\mathbf{X}|$ is the length of the sequence. Denote intensity labels associated with $\mathbf{X}$ as $\mathbf{Y} = \{y_i \in \mathbb{R} | i \in \mathbf{V}\}$, where $\mathbf{V} \subseteq \{1, ..., |\mathbf{X}|\}$ is a subset of indices for the sequence. We are interested in exploring different settings on $\mathbf{V}$. For a fully supervised problem, $\mathbf{V}$ contains all the frame indices. For a weakly supervised problem, $\mathbf{V}$ only contains selected frame indices. Finally, for a fully unsupervised problem, $\mathbf{V}$ is an empty set.

We assume that within a sequence, the expression intensity either increases monotonically until it reaches its peak or decreases monotonically after passing its peak, where peak is attained at apex frames. Specifically, let p be the index of apex frame, then the following inequalities hold

$$y_i \geq y_j, \ \forall (i,j) \in \mathbf{E} \tag{1}$$

where $\mathbf{E} = \{(i,j) | 1 \leq j < i \leq p \text{ or } p \leq i < j \leq |\mathbf{X}|\}$ is a set specifying pairwise ordinal relationship. Intuitively speaking, the expression can only evolve in an onset-apex-offset fashion.

During training, we are provided with multiple sequences and additional information on intensity annotations $\mathcal{D} = \{\mathbf{X}_n, \mathbf{Y}_n, \mathbf{V}_n, \mathbf{E}_n\}, n = 1, ..., N$, where N is the number of sequences. Intensity label set $\mathbf{V}_n$ may vary depending on learning settings. In fully supervised setting, $\mathbf{V}_n = \{1, ..., |\mathbf{X}_n|\}$ *i.e.* all the frames. In weakly supervised setting, $\mathbf{V}_n = \{1, p, |\mathbf{X}_n|\}$ *i.e.* onset, apex and offset frames, assuming the first and the last frame are onset and offset frame respectively. In unsupervised setting, $\mathbf{V}_n = \emptyset$. In each setting, $\mathbf{E}_n$ is available given $\mathbf{V}_n$.

Our goal is to learn a regression function $f : \mathbb{R}^d \mapsto \mathbb{R}$ applied on frame-level under different learning settings of $\mathbf{V}_n$. During evaluation, given an image of expression, we perform expression intensity estimation as

$$y = f(\mathbf{x}; \theta) \tag{2}$$

where θ is the parameter of the function f and y is the ground truth expression intensity. For dataset without intensity annotation, we use relative intensity as substitution. We will define relative intensity specifically later in section 5. We consider different types of expressions separately.

3. Background

We briefly review Support Vector Regression (SVR) and Ordinal Regression (OR), on which our method is based.

3.1. Support Vector Regression

As a modification of Support Vector Machine (SVM) [4], SVR [27] learns a regression model given data-label pair $\{y_i, \mathbf{x}_i\}$. One of the most commonly used variant called ϵ-insensitive SVR learns the model parameter $\theta = \{\mathbf{w}, b\}$ by solving the following optimization problem.

$$\begin{aligned} \min_{\theta,\eta^+,\eta^-} \ & \frac{1}{2}||\mathbf{w}||^2 + \gamma \sum_i (\eta_i^+ + \eta_i^-) \\ \text{s.t. } & \mathbf{w}^T\phi(\mathbf{x}_i) + b - y_i \leq \epsilon + \eta_i^+ \\ & y_i - \mathbf{w}^T\phi(\mathbf{x}_i) - b \leq \epsilon + \eta_i^- \\ & \eta_i^+, \eta_i^- \geq 0, \ \forall i \end{aligned} \tag{3}$$

where ϵ is a constant which defines the maximum deviation allowed for a prediction to be considered as correct, $\phi : \mathcal{X} \mapsto \mathcal{F}$ is a mapping from input space $\mathcal{X}$ to some feature space $\mathcal{F}$ and γ is a constant balancing between the regularization and regression loss. Solution to Eq.(3) can be obtained by solving its dual problem, which avoids explicit computation of $\phi(\mathbf{x}_i)$ using kernel trick [28].

An important property inherited from SVM is that the solution of SVR is sparse in the sense that the model parameter can be determined using only a subset of data points namely the support vectors. For further details, readers are refered to [28]. Since SVR uses the label information associated with each data point, it can only be trained using frames with known intensity labels. We will use SVR as our baseline method for comparison.

3.2. Ordinal Regression

In OR, the target value of each data point is an ordinal variable, which gives a ranking among different data points. It is widely used in information retrieval task such as ranking data according to their relevance to the query [16]. In such scenario, the ordinal relationship rather than the absolute regression value is of primary interest. We introduce the formulation proposed by Herbrich *et al.* [10]. Given data-label pair $\{y_i, \mathbf{x}_i\}$, where y_i is a discrete ordinal variable, the model parameter $\mathbf{w}$ is learned by solving the following optimization problem.

$$\begin{aligned} \min_{\mathbf{w},\xi} \ & \frac{1}{2}||\mathbf{w}||^2 + \gamma \sum_{(i,j)\in E} \xi_{ij} \\ \text{s.t. } & \mathbf{w}^T(\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)) \geq 1 - \xi_{ij} \\ & \xi_{ij} \geq 0, \ \forall (i,j) \in \mathbf{E} \end{aligned} \tag{4}$$

where $\mathbf{E} = \{(i,j)|y_i \geq y_j\}$ is the ordinal set. Eq.(4) tries to find a regression function $f(\mathbf{x};\theta)$ which minimizes the number of swapped pairs in training data. In general, the number of constraints is $O(n^2)$ where n is the total number of data points, which can be problematic for large scale problems. More recently, Joachims [12] proposed a formulation which condenses the number of constraints used in Eq.(4) and solved it using cutting-plane algorithm. The complexity is reduced to $O(n^d),\ d < 2$.

Different from SVR, OR does not need the intensity annotations of any frames in a sequence except the ordinal position of the frame which comes with the sequence itself. We consider this as an unsupervised approach and will use OR as our second baseline method for comparison.

4. Proposed Method

4.1. Ordinal Support Vector Regression

Our model is motivated by the two baseline methods which sit at the two ends of regression model learning. SVR only uses intensity labels of annotated frames while ignoring the temporal order and OR only focuses on ordinal relationship without using labeled intensity values. We propose a max-margin based regression model which incorporates the benefits of both models by taking advantage of some labeled frames and readily available ordinal information comes with the sequence that satisfying the assumptions stated in section 2. We also develop an efficient algorithm to solve the optimization problem for model learning.

We use linear model $f(\mathbf{x};\theta) = \mathbf{w}^T\mathbf{x} + b$ with model parameter $\theta = \{\mathbf{w}, b\}$. Given $\mathcal{D} = \{\mathbf{X}_n, \mathbf{Y}_n, \mathbf{V}_n, \mathbf{E}_n\}, n = 1, ...N$, we solve the following optimization problem.

$$\begin{aligned} \min_{\theta,\eta,\xi} \ & \frac{1}{2}||\mathbf{w}||^2 + \gamma_1 \sum_{n=1}^{N} \sum_{k\in\mathbf{V}_n} (l_1(\eta_k^{(n)+}) + l_1(\eta_k^{(n)-})) \\ & + \gamma_2 \sum_{n=1}^{N} \sum_{(i,j)\in\mathbf{E}_n} l_2(\xi_{ij}^{(n)}) \\ \text{s.t. } & \mathbf{w}^T\mathbf{x}_k^{(n)} + b - y_k^{(n)} \leq \epsilon + \eta_k^{(n)+} \\ & y_k^{(n)} - \mathbf{w}^T\mathbf{x}_k^{(n)} - b \leq \epsilon + \eta_k^{(n)-} \\ & \mathbf{w}^T(\mathbf{x}_i^{(n)} - \mathbf{x}_j^{(n)}) \geq 1 - \alpha_{ij}\xi_{ij}^{(n)} \\ & \eta_k^{(n)+}, \eta_k^{(n)-}, \xi_{ij}^{(n)} \geq 0 \\ & \forall k \in \mathbf{V}_n, (i,j) \in \mathbf{E}_n, \ n = 1, ..., N \end{aligned} \tag{5}$$

where $\gamma_1, \gamma_2, \epsilon > 0$ are constants. l_1 and l_2 are some functions applied on slack variables η and ξ respectively. $\mathbf{V}_n$ and $\mathbf{E}_n$ are the intensity label set and ordinal set associated with the n^{th} sequence as defined in section 2.

The first two sets of constraints are adopted from SVR in order to restrict regression function to fit the provided intensity labels. The third set of constraints is adapted from OR in order to take advantage of the temporal evolution pattern of the sequence. The additional parameter $\alpha = \{\alpha_{ij}\} = \{\frac{1}{|i-j|}\}, \forall i \neq j$ is introduced to encourage temporal smoothness, namely similar feature values between temporally close frames. Noticing that setting all $\alpha_{ij} = 1$ reduces to the same constraints used in OR.

Intuitively, Eq.(5) tries to find a regression function which balances the regression loss and ordinal loss at the same time. More importantly, this formulation is flexible in handling different cases of label annotations *i.e.* fully annotated, partially annotated and un-annotated. Most of previous work has been focusing on either fully supervised approach [25, 13, 24] or unsupervised approach [31], our method can in addition to handle a weakly supervised setting where only selected key frames are annotated with intensity. To the best of our knowledge, this is the first work that addresses continuous expression intensity estimation under a variety of annotation settings. We refer to this method as OSVR.

For the choices of function l_1 and l_2, we consider two different configurations. The first one we set $l_i(x) = x$, which gives us hinge loss on both regression and ordinal constraints. The second one we set $l_i(x) = x^2$, which gives us squared hinge loss. However, the formulation is completely general and can be extended to apply different choices of loss functions. Noticing parameters γ_1 and γ_2 effectively balance two types of losses. By setting extreme large value to either γ_1 or γ_2 will forces regression loss dominates ordinal loss or vice versa. Their values can be determined by cross-validation in practice.

4.2. Alternating Direction Method of Multipliers

The number of constraints in Eq.(5) is $O(n^2)$, where n is the number of frames. In our experiment, n is usually at the order of $10^2 \sim 10^3$, resulting the number of constraint $10^4 \sim 10^6$. We resort to ADMM due to its compactness in handling optimization problem with large number of constraints. The use of augmented Lagrangian multipliers further improves the efficiency in terms of fast convergence rate. Adopting the notation used in [1], we consider the minimization problem with respect to variables $u \in \mathbb{R}^n, z \in \mathbb{R}^m$.

$$\min_{u,z} f(u) + g(z) \quad \text{s.t. } Au + Bz = c \tag{6}$$

where f and g are some convex functions, $A \in \mathbb{R}^{p \times n}, B \in \mathbb{R}^{p \times m}, c \in \mathbb{R}^p$ are matrix or vector with proper dimension, the augmented Lagrangian can be written as follows

$$L_\rho(u, z, v) = f(u) + g(z) + v^T(Au + Bz - c) + \rho/2||Au + Bz - c||_2^2 \tag{7}$$

where $v \in \mathbb{R}^p$ is the ordinary Lagrangian multipliers corresponding to p equality constraints. $\rho > 0$ is the augmented Lagrangian multiplier. ADMM solves the optimization problem by iteratively updating variables as follows

$$u^{k+1} := \arg\min_u L_\rho(u, z^k, v^k) \tag{8}$$

$$z^{k+1} := \arg\min_z L_\rho(u^{k+1}, z, v^k) \tag{9}$$

$$v^{k+1} := v^k + \rho(Au^{k+1} + Bz^{k+1} - c) \tag{10}$$

Since f and g are convex functions, L_ρ is also convex. Therefore, in general we can find unique optimal solution in Eq.(8) and Eq.(9). The convergence of these updates can be proved under fairly mild assumptions [1], where neither f nor g has to be differentiable.

4.3. Solving OSVR using ADMM

We now reformulate Eq.(5) and derive the needed updates in order to apply ADMM. Define variable $u = \theta \equiv [\mathbf{w}; b] \in \mathbb{R}^{d+1}, z \in \mathbb{R}^{M_1+M_2}$, where d is the feature dimension. M_1 and M_2 are the total number of constraints corresponding to regression loss (the first two sets of constraints in Eq.(5)) and ordinal loss (the third set of constraints in Eq.(5)). Define an element-wise operator $\lfloor\cdot\rfloor_0$ which truncates negative value to 0. We can rewrite Eq.(5) as

$$\min_{u,z} \frac{1}{2} u^T \Lambda u + \mu^T l(\lfloor z \rfloor_0) \quad \text{s.t. } Au + c = z \tag{11}$$

where l is selected loss function and $\Lambda \in \mathbb{R}^{(d+1)\times(d+1)}$ is a diagonal matrix. $\mu \in \mathbb{R}^{M_1+M_2}$ is a vector whose first M_1 entries are γ_1 and the last M_2 entries are $\gamma_2\alpha$. $A \in \mathbb{R}^{(M_1+M_2)\times(d+1)}$ is a matrix and $c \in \mathbb{R}^{M_1+M_2}$ is a vector whose values are chosen in the way to resemble the first three sets of linear constraints in Eq.(5) with inequality replaced by equality. Specifically,

$$A = \begin{bmatrix} \mathbf{X}_V & \mathbf{1} \\ -\mathbf{X}_V & -\mathbf{1} \\ -\mathbf{X}_E & \mathbf{0} \end{bmatrix}, \; c = \begin{bmatrix} -\epsilon\mathbf{1} - \mathbf{y} \\ -\epsilon\mathbf{1} + \mathbf{y} \\ \mathbf{1} \end{bmatrix}$$

where $\mathbf{X}_V$ is a matrix whose rows are data samples with known intensity labels and $\mathbf{X}_E$ is a matrix whose rows are the difference between two data samples whose frame indices belong to the ordinal set. $\mathbf{y}$ is a vector of known intensity labels. $\mathbf{1}$ and $\mathbf{0}$ are vectors with proper dimension containing all 1s and 0s respectively.

Noticing that z is not restricted to be positive in Eq.(11), which allows the equality to be attained. However, the objective function remains unaffected due to the negative truncation operator on z. In other words, $\lfloor z \rfloor_0$ correspond to the non-negative slack variables η^+, η^-, ξ introduced in Eq.(5). u by definition gives the values of parameter θ. Therefore, the optimal solution to Eq.(11) is the same as that to Eq.(5).

The augmented Lagrangian has quadratic form with respect to u, z and is linear to v. Even though we have non-differentiable term $\lfloor z \rfloor_0$, we can still compute a simple close-form solution with a compact soft thresholding operator [1]. We provide the detailed derivation in the supplementary material and only list the main results here. In the hinge loss case, the updates corresponding to Eq.(8)-Eq.(10) are given by

$$u^{k+1} := [\frac{1}{\rho}\Lambda + A^T A]^{-1} A^T (z^k - \frac{1}{\rho} v^k - c) \tag{12}$$

$$z_i^{k+1} := S_{\frac{\mu_i}{2\rho}}(a_i) \tag{13}$$

$$v^{k+1} := v^k + \rho(Au^{k+1} - z^{k+1} + c) \tag{14}$$

where $a = \frac{1}{\rho} v^k + Au^{k+1} + c - \frac{1}{2\rho}\mu \equiv \{a_i\}$ and

$$S_\kappa(a_i) = \begin{cases} a_i - \kappa, & \text{if } a_i > \kappa \\ 0, & \text{if } |a_i| \leq \kappa \\ a_i + \kappa, & \text{if } a_i < -\kappa \end{cases} \tag{15}$$

is the soft thresholding operator for some constant $\kappa > 0$ and the subscript i is referring to the i^{th} entry in each vector.

For the squared hinge loss case, updates on u and v remain the same. Update on z is given by

$$z_i^{k+1} := \begin{cases} \frac{\rho a_i}{\rho + 2\mu_i}, & \text{if } a_i \geq 0 \\ a_i, & \text{if } a_i < 0 \end{cases} \tag{16}$$

where $a = \frac{1}{\rho} v^k + Au^{k+1} + c \equiv \{a_i\}$. As the algorithm converges to the optimal solution we have

$$z^{k+1} - z^k \to \mathbf{0} \tag{17}$$

$$Au^{k+1} - z^{k+1} + c \to \mathbf{0} \tag{18}$$

We stop the updates if the LHS of both Eq.(17) and Eq.(18) become smaller than some tolerance value or reaches the maximum number of iterations. The overall process of ADMM is summarized in Algorithm 1[1]. After solving the optimization problem, we can predict the intensity given new frame $\mathbf{x}'$ as $y' = \mathbf{w}^T \mathbf{x}' + b$.

Algorithm 1 OSVR learning by ADMM

Input: $\mathbf{X}$: expression sequences, $\mathbf{Y}$: intensity values, $\mathbf{V}$: intensity label set, $\mathbf{E}$: ordinal set

Output: Model parameters

Construct Λ, μ, A, c using $\mathbf{X}, \mathbf{Y}, \mathbf{V}, \mathbf{E}$
$u \leftarrow \mathbf{0}$, $z \leftarrow \mathbf{0}$, $v \leftarrow \mathbf{0}$
repeat
  update u using Eq.(12)
  update z using Eq.(13) or Eq.(16)
  update v using Eq.(14)
until convergence or reach maximum iteration number
return u

5. Experiment

We perform intensity estimation under different learning settings depending on the use of ground truth intensity labels. For dataset with complete frame-level intensity annotation, we learn models under fully supervised, weakly supervised and unsupervised settings and evaluate them against ground truth intensities (GTI). For dataset without frame-level intensity annotation, we introduce ordinal relationship based relative intensity (RI) as substitution to ground truth. Our models are learned under weakly supervised setting and are evaluated against RI. The overall experiment process is shown in Figure 2. The specific setting for each dataset is listed in Table 1.

For each frame, we extract three types of features namely facial landmark points, local binary pattern (LBP) [20] and Gabor wavelet coefficients [19]. Specifically, we use IntraFace [30] to track 66 facial landmark points. Then each frame is aligned by performing an affine transformation such that landmark points connecting two eyes are horizontal. We crop the face location and resize it to 100×100 pixels, from where the LBP and Gabor features are extracted. We choose uniform LBP with 8-neighbourhood pixels. The image is divided into 5 equally sized non-overlapping patches. LBP histograms are extracted from each patch, resulting 1475-dimensional vectors. Gabor features are extracted from the same patches, yielding 1000-dimensional vectors. We apply PCA to each type of features separately to keep up to 95% energy. Final feature vector concatenates PCA results of each set of feature.

Table 1. Different learning and evaluation settings.

Learning		Evaluation		
Setting	Intensity label	PAIN	CK+	BU-4DFE
Fully supervised	All frames GTI	All frames GTI	All frames GTI	All frames GTI
Weakly supervised	Key frames GTI			
Unsupervised	None		RI	RI

5.1. Datasets

We select three datasets in our experiment. Some sample sequences from each dataset are shown in Figure 1.

UNBC-McMaster shoulder pain [18] (PAIN) dataset captures expressions from subjects suffering from shoulder pain. This dataset contains 200 spontaneous expression sequences, and the sequences are FACS coded frame-by-frame. The dataset provides pain intensity value calculated using the Prkachin and Solomon pain intensity (PSPI) met-

[1]Code available from http://bit.ly/OrdinalSVR

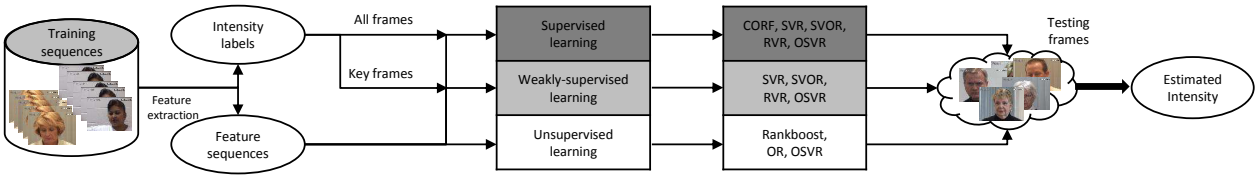

Figure 2. A diagram showing the experiment process. Depending on the experiment setting, different amounts of intensity annotation information are fed into model learning process, resulting different models. Training is performed using complete expression sequences while testing is performed on each frame of a sequence.

ric [21] for every frame. We consider this as ground truth intensity.

Extended CK [17] (CK+) dataset contains 593 posed expression sequences from 123 subjects aged from 18 to 30 years old. Each sequence begins with an onset frame and ends at an apex frame. No intensity annotation is provided. Subjects are requested to perform 7 basic categories of expressions: anger, contempt, disgust, fear, happy, sadness and surprise.

BU-4DFE [32] dataset records 606 3D dynamic facial sequences (called 4D data) from 101 subjects aged from 18 to 45 years old. Subjects are requested to perform six expressions including anger, disgust, fear, happy, sadness and surprise. Each sequence starts and ends with neutral face and the expression evolves by onset-apex-offset temporal pattern. No intensity annotation is provided.

5.2. Experiment on PAIN dataset

We select 191 out of 200 sequences with total number of frames 6497, excluding one subject whose expressions do not have noticeable pain. The average length of selected sequences is about 240. Each frame is labeled with discrete intensity level from 0 to 15. The vast majority ($\approx 81.6\%$) of the data contains no noticeable pain where the intensity is labeled as 0. On the other hand, severe pain (level$\geq$6) is also rare ($\approx 1.4\%$). In order to make a dataset with more balanced intensity levels, we perform the following pre-processing on each sequence. First, we perform additional quantization on the intensity level by aggregating different intensity levels in the same way as [23]. Second, we perform an adaptive down-sampling on the entire sequence. If the intensity level remains the same for more than 5 consecutive frames, we choose the first one as representative frame. We find this strategy is effective in both preserving intensity pattern and reducing redundant samples especially the ones with intensity level 0. Finally, we segment the sequences into subsequences where the starting and ending frame has intensity level 0. The final distribution of intensity levels is listed in Table 2.

During training phase, we randomly select 60% of sequences as training set and remaining as validation set for the purpose of selecting constant γ_1, γ_2. We fix the value of $\epsilon = 0.1$ and $\rho = 0.1$, which we find insensitive to the final result. During testing phase, we perform leave-one-subject-out test and the results are averaged over all subjects.

Table 2. Aggregated pain intensity level and the portion of each level over total number of frames

Quantized	Annotated	Pain description	Portion (%)
0	0	None	68.5
1	1	Mild	10.7
2	2	Discomforting	8.9
3	3	Distressing	5.6
4	$4 \sim 5$	Intense	4.1
5	$6 \sim 15$	Excruciating	2.2

Table 3. Results of different methods on PAIN dataset.

Setting	Method	PCC	ICC	MAE
Fully supervised	KCORF$_h$[23]	N/A	*0.7030*	*0.8000*
	csCORF$_{wh}$[24]	N/A	*0.6400*	*0.8200*
	SVR [27]	0.5659	0.5045	0.8538
	SVOR [3]	0.5483	0.3726	0.9366
	RVR[13]	0.5749	0.5036	0.8687
	OSVR-L1	0.5999	**0.5593**	1.0252
	OSVR-L2	**0.6014**	0.5335	**0.8095**
Weakly supervised	SVR [27]	0.4766	0.4511	1.3895
	SVOR [3]	0.5051	0.4240	2.9801
	RVR[13]	0.4823	0.4365	1.1122
	OSVR-L1	0.4981	0.4710	1.1512
	OSVR-L2	**0.5441**	**0.4955**	**0.9519**
Unsupervised	Rankboost[31]	0.4341	0.3718	**1.0609**
	OR [10]	0.4572	**0.4279**	2.0903
	OSVR-L1	0.4921	0.4020	2.6399
	OSVR-L2	**0.5101**	0.4108	1.1180

We use Pearson correlation coefficient (PCC), intra-class correlation (ICC) and mean absolute error (MAE) as evaluation metrics. PCC measures how well the prediction can capture the trend of intensity change. We use ICC(3,1) as defined in [26] to measure the consistency within each intensity level. Both PCC and ICC are numbers between 0 and 1 and the larger the better. MAE is a positive number measuring the deviation between prediction and actual value and the smaller the better. We compare our method with RVR [13], SVOR [3], KCORF [23], csCORF [24] and Rankboost [31]. The baseline methods SVR and OR are im-

Figure 3. Predicted intensity and annotated intensity on selected consecutive frames for different methods under different annotation settings. Left: Fully supervised setting. Middle: Weakly supervised setting. Right: Unsupervised setting. (Best view in color)

plemented by liblinear [7] and rank-SVM [12] respectively. The results are shown in Table 3.

In fully supervised setting, both OSVR-L1 and OSVR-L2 achieve higher PCC and ICC comparing to SVR, RVR and SVOR. This shows that using additional ordinal information help increase the fitting of the trend of intensity change. The MAE values of OSVR are net effect of two factors. On one hand, the use of intensity label for each frame diminishes the effect of ordinal information and the intensity level distribution is uneven with a large portion of low intensity values. On the other hand, we reinforce ordinal information by encouraging temporal smoothness. Overall, OSVR-L2 achieves the best MAE result. For KCORF, we list the original results reported in [23] where fixed testing set instead of leave-one-subject-out test is used for experiment. Similarly, original results of csCORF in [24] are listed. Since the same intensity level annotation is used, we list the results for completeness despite that they are not directly comparable.

In a weakly supervised setting, on average, 3 out of 34 frame labels per sequence are used for learning. In other words, only 8.8% of total label information is required in such setting. Two variants of OSVR both show noticeable improvement comparing to competing methods. The results under unsupervised setting indicate OSVR are comparable to others. This is because without any intensity label, the regression loss is essentially defective. In addition, some sequences are labeled as 0 for all the frames, which makes apex and onset frames indifferent and thus reduces the effectiveness of pure ordinal information based approach.

Comparing OSVR learned under different settings, the results are better with more label information available. The advantage of OSVR lies in exploiting absolute intensity labels (if available) and ordinal information. It can adapt to fully supervised and unsupervised settings. In particular, OSVR demonstrates superiority especially under weakly supervised setting. Figure 3 shows the predicted values by different methods under different settings of some consecutive frames versus ground truth.

5.3. Experiment on CK+ and BU-4DFE datasets

For CK+ dataset, we select 327 sequences from 118 subjects, excluding 266 sequences without expression labels. Total number of selected frames is 5876. The first and last frame of each sequence are onset and apex frame respectively. For BU-4DFE dataset, we select 120 sequences from 20 subjects and manually identify the apex frame for each sequence, yielding 2289 frames in total.

For both datasets, we define the minimum and maximum relative intensity of each sequence as $I_l = 0$ and $I_h = 10$ respectively. In addition, I_l is attained at onset or offset frame. I_h is attained at apex frame. I_l and I_h are the same for different sequences so that we can compare intensity values across different subjects. To evaluate the performance of intensity estimation, we artificially assign the intensity label y_j of a frame using its relative distance to the corresponding apex frame p of the sequence using

$$y_j = \frac{j-1}{p-1}(I_h - I_l)\delta_{j<p} + \frac{m-j}{m-p}(I_h - I_l)\delta_{j\geq p} \quad (19)$$

where $j = 1, ..., m$ and m is the length of the sequence. δ is the indicator function. Eq.(19) essentially produces intensity curve change linearly between onset/offset and apex frames. We perform intensity estimation separately for each type of emotion. One limitation of such relative intensity is ignoring the within class variation of expression intensity. We use the same evaluation metrics, namely PCC, ICC and MAE. For CK+, we perform 10-fold cross subjects test. For BU-4DFE, we perform leave-one-subject-out test. The results are computed using all the testing frames for each type of emotion. For comparison, we use SVR[27], RVR[13], SVOR[3], GPOR[2], Rankboost[31] and OR[10]. All the methods except Rankboost and OR are trained given relative intensity of key frames. Rankboost and OR are trained using ordinal information only. For evaluation, relative intensity of testing frames are used for all the methods. The results of two datasets are shown in Figure 4. We use bar graph for compactness, the exact values of each method are listed in the supplementary material.

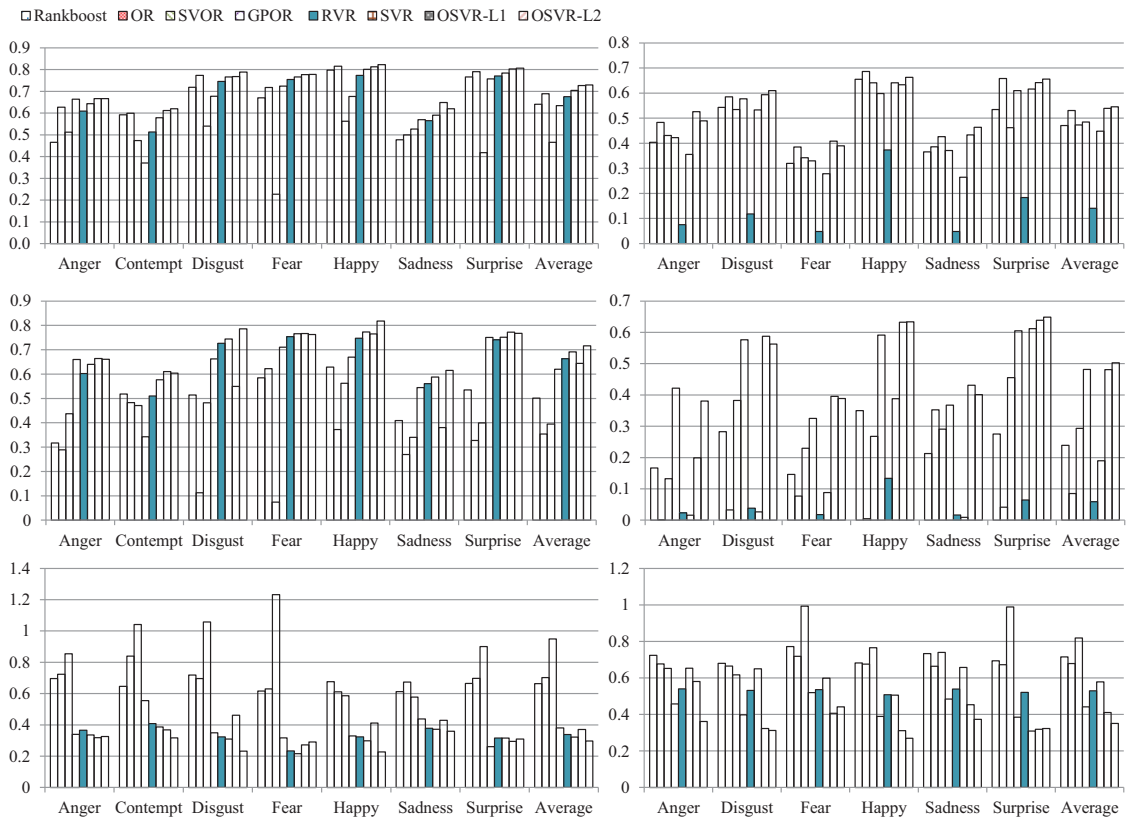

Figure 4. Results on CK+ (left column) and BU-4DFE (right column) datasets using different methods for each emotion type. From top to bottom are PCC, ICC and log(MAE) values (Best view in color). The exact values of the results are listed in the supplementary material.

In experiment, we found setting temporal smoothness coefficients $\alpha = 1$ yield slightly better results. Therefore, we report the results obtained with $\alpha = 1$. For CK+ dataset, one variant of OSVR outperforms the competing methods on all three metrics. In particular, averaging over different expressions, OSVR-L2 improves PCC, ICC and MAE by 3.5%, 3.6% and 5.7% respectively comparing to the best competing method SVR. For BU-4DFE dataset, both variants of OSVR are better than or equal to competing methods on all metrics. Comparing to the second best method, OSVR-L2 improves PCC, ICC and MAE by 2.8%, 4.4% and 33.8% respectively. OSVR-L1 improves PCC and MAE by 1.7% and 24.0% respectively. Although the improvements on PCC and ICC are comparable to CK+, the improvement on MAE is substantial. Considering BU-4DFE as a more challenging case where the expression intensity changes in different ways before and after apex frames, the proposed method achieves both higher PCC, ICC and lower MAE at the same time.

6. Conclusion and Future Work

In this work, we formalized a regression problem for frame-level expression intensity estimation. By exploiting the intrinsic ordinal relationship among different frames in an expression sequence, the model learning is compatible with different levels of supervision on expression intensity annotation. This is very useful in case intensity annotation is not available for all the frames. In particular, we use key frames including onset, apex and offset as weak supervision on regression model learning. Results on three different benchmark datasets with different intensity annotation information show that the proposed OSVR method outperforms existing approaches supplied with only key frames' annotation. OSVR can adapt to fully supervised and unsupervised learning setting with comparable performance to other methods. We consider several extensions including introducing kernel to proposed method and adding dynamic information such as expression transition as additional constraints. More sophisticated temporal smoothness scheme can be introduced. We also plan to extend current framework to AU intensity estimation.

Acknowledgment

This paper was supported in part by a contract from the General Electric and by the National Science Foundation of China (Grant No. 61473270). We thank Dr. Kristin P. Bennett for the discussion on optimization methods.

References

[1] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011. 4, 5

[2] W. Chu and Z. Ghahramani. Gaussian processes for ordinal regression. In *Journal of Machine Learning Research*, pages 1019–1041, 2005. 7

[3] W. Chu and S. S. Keerthi. New approaches to support vector ordinal regression. In *ICML*, pages 145–152. ACM, 2005. 6, 7

[4] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. 3

[5] A. Dhall and R. Goecke. Group expression intensity estimation in videos via gaussian processes. In *ICPR*, pages 3525–3528. IEEE, 2012. 1

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

[7] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008. 7

[8] B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. *Pattern recognition*, 36(1):259–275, 2003. 1

[9] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars. Modeling video evolution for action recognition. In *CVPR*, volume 2, page 8, 2015. 2

[10] R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. *Advances in neural information processing systems*, pages 115–132, 1999. 3, 6, 7

[11] U. Hess, S. Blairy, and R. E. Kleck. The intensity of emotional facial expressions and decoding accuracy. *Journal of Nonverbal Behavior*, 21(4):241–257, 1997. 1, 2

[12] T. Joachims. Training linear svms in linear time. In *SIGKDD*, pages 217–226. ACM, 2006. 3, 7

[13] S. Kaltwang, O. Rudovic, and M. Pantic. Continuous pain intensity estimation from facial expressions. In *Advances in Visual Computing*, pages 368–377. Springer, 2012. 1, 2, 4, 6, 7

[14] K. K. Lee and Y. Xu. Real-time estimation of facial expression intensity. In *ICRA*, volume 2, pages 2567–2572. IEEE, 2003. 2

[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. In *FG*, pages 1–7. IEEE, 2013. 1

[16] T.-Y. Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009. 3

[17] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *CVPRW*, pages 94–101. IEEE, 2010. 2, 6

[18] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *FG*, pages 57–64. IEEE, 2011. 1, 2, 5

[19] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. In *FG*, pages 200–205. IEEE, 1998. 5

[20] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, 2002. 5

[21] K. M. Prkachin and P. E. Solomon. The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain. *Pain*, 139(2):267–274, 2008. 1, 6

[22] O. Rudovic, V. Pavlovic, and M. Pantic. Multi-output laplacian dynamic ordinal regression for facial expression recognition and intensity estimation. In *CVPR*, pages 2634–2641. IEEE, 2012. 2

[23] O. Rudovic, V. Pavlovic, and M. Pantic. Automatic pain intensity estimation with heteroscedastic conditional ordinal random fields. In *Advances in Visual Computing*, pages 234–243. Springer, 2013. 2, 6, 7

[24] O. Rudovic, V. Pavlovic, and M. Pantic. Context-sensitive dynamic ordinal regression for intensity estimation of facial action units. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(5):944–958, 2015. 1, 2, 4, 6, 7

[25] A. Savran, B. Sankur, and M. T. Bilge. Regression-based intensity estimation of facial action units. *Image and Vision Computing*, 30(10):774–784, 2012. 1, 4

[26] P. E. Shrout and J. L. Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979. 6

[27] A. Smola and V. Vapnik. Support vector regression machines. *Advances in neural information processing systems*, 9:155–161, 1997. 3, 6, 7

[28] A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004. 3

[29] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan. Toward practical smile detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(11):2106–2111, 2009. 2

[30] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, pages 532–539. IEEE, 2013. 5

[31] P. Yang, Q. Liu, and D. N. Metaxas. Rankboost with l1 regularization for facial expression recognition and intensity estimation. In *CVPR*, pages 1018–1025. IEEE, 2009. 2, 4, 6, 7

[32] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale. A high-resolution 3d dynamic facial expression database. In *FG*, pages 1–6. IEEE, 2008. 2, 6