Local-Global Ranking for Facial Expression Intensity Estimation

Tadas Baltrušaitis
Language Technologies Institute
Carnegie Mellon University
Email: tbaltrus@cs.cmu.edu

Liandong Li
Beijing Normal University
Email: bnulee@hotmail.com

Louis-Philippe Morency
Language Technologies Institute
Carnegie Mellon University
Email: morency@cs.cmu.edu

***Abstract*—Facial action units provide an objective characterization of facial muscle movements. Automatic estimation of facial action unit intensities is a challenging problem given individual differences in neutral face appearances and the need to generalize across different pose, illumination and datasets. In this paper, we introduce the Local-Global Ranking method as a novel alternative to direct prediction of facial action unit intensities. Our method takes advantage of the additional information present in videos and image collections of the same person (e.g. a photo album). Instead of trying to estimate facial expression intensities independently for each image, our proposed method performs a two-stage ranking: a local pair-wise ranking followed by a global ranking. The local ranking is designed to be accurate and robust by making a simple 3-class comparison (higher, equal, or lower) between randomly sampled pairs of images. We use a Bayesian model to integrate all these pair-wise rankings and construct a global ranking. Our Local-Global Ranking method shows state-of-the-art performance on two publicly-available datasets. Our cross-dataset experiments also show better generalizability.**

1. Introduction

Over the past few years, there has been an increased interest in machine understanding and recognition of affective and cognitive mental states, especially based on facial expression analysis [23]. The face is one of the main channels of nonverbal communication, automatic facial expression analysis can be used to facilitate human computer interaction [4], [21] and to better understand mental illness [11], [29]. facial action units [8] (AUs) provide a way to objectively describe facial expressions through the movement of individual or groups of facial muscles. Automatic AU analysis is concerned with detecting AUs as they occur on the face and estimating their intensities. The latter is especially important as it allows us to analyze expression dynamics in videos and to have a more fine-grained understanding of a person's expressivity.

While facial expression analysis has made huge progress [18] there are still a lot of challenges facing automatic estimation of AU intensity. One such challenge is the idiosyncrasy of faces, some faces look more *smiley* or *frowny* even when the facial expression is neutral. Accounting for such individual differences is difficult and few approaches consider them. The other big issue facing AU intensity estimation is the common inability of predictive models to generalize well across datasets due to changes in pose, illumination, and even ground truth reliability. Finally, recognizing subtle variations in AU intensity is very difficult without seeing the range of expressions. These difficulties are especially apparent when performing direct AU recognition – based on a classification of a single face image; see Figure 1 for an illustration.

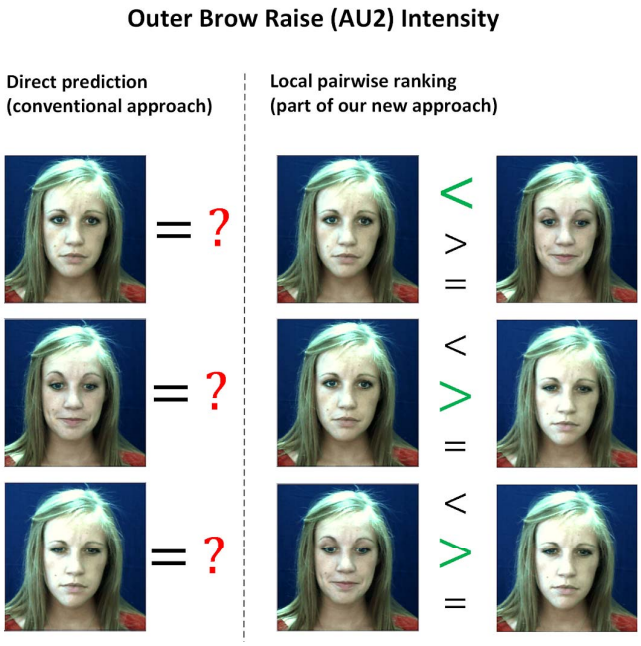

Figure 1: The typical approach for facial action unit intensity estimation most often relies on direct prediction from an individual image. Our Local-Global Ranking model approaches the task through an easier sub-task – local pairwise image comparisons. Our model is able to handle personal idiosyncrasies such as higher or more accentuated eyebrows robustly.

In this paper, we introduce a novel Local-Global Ranking method which takes advantage of the extra information present in videos and image collections of the same person

978-1-5386-0563-9/17/$31.00 ©2017 IEEE

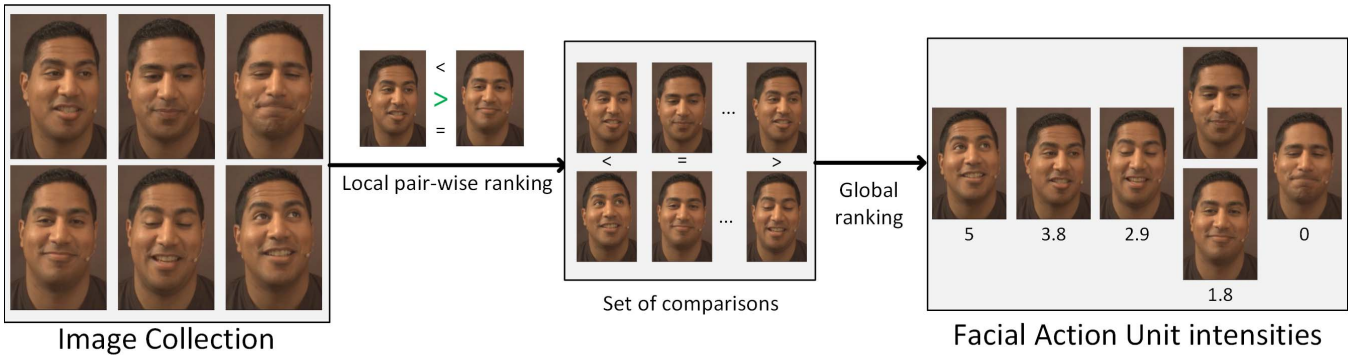

Figure 2: The outline of our approach for AU intensity estimation. We first randomly sample pairs of images from the image collection. For each image pair, we perform a local pairwise ranking (see Section 3.1). This leads to a set of local comparisons $\mathbf{r}$ (see Section 3.1.1). We then use a global Bayesian ranking approach to recover the AU intensity ranks (see Section 3.2) which are converted to actual intensities using a transformation ϕ (Equation 8).

(e.g. a photo album) showing a range of expressions. Instead of estimating facial expression intensities directly and independently for each image, we perform two-stage ranking: local ranking and global ranking. The overview of our approach can be seen in Figure 2. We first classify randomly sampled pairs of images from the image collection and then classify them as being of higher/lower/same intensity. We then use a Bayesian model to integrate all of these pairwise ranking to construct a global ranking. Our approach is person-independent, but its strength lies in learning to notice expression differences within a person. We believe it is an easier problem than the conventional approach of classifying expressions directly using a single image. This intuition is supported by our experimental results and the fact that humans find the task of ranking images to be easier than giving absolute values to them [26].

Our Local-Global Ranking approach has a number of advantages over conventional ways for performing facial action unit intensity estimation: 1) it leads to an improvement in performance over direct prediction; 2) it achieves better cross-dataset generalizeability; 3) by controlling for a number of local comparisons our approach has a speed/accuracy trade-off allowing for real-time and offline processing.

2. Background

Facial action unit (AU) recognition has received a lot of attention over the past years [7], [18], [25], [30]. A number of approaches have been proposed to estimate their presence and intensity. We provide a short survey of the latter, we also provide a brief discussion on ranking-based learning.

A lot of work on AU intensity recognition focused on directly predicting AU intensity values from a single image [2], [22], [30], [31]. Some examples of such approaches include using facial geometry and Local Binary Gabor Patterns [32] extracted from a single face image to predict AU intensities using a Support Vector Machine [30]. Wang et al. [31] proposed a Bayesian Network that captures the relationship between facial expressions and AUs. Kaltwang et al. [15] demonstrated how to use a generative latent tree model to address the uncertainty of input features.

Other work includes Baltrušaitis et al. [1], who proposed to use a continuous output extension of Conditional Random Fields - Continuous Conditional Neural Fields (CCNF) to model non-linear and temporal relationships between facial appearance and AU intensities. Rudovic et al. [24] proposed to model context in which the AUs occur and to employ ordinal regression to account for non-linear relationship between different AU intensity levels. More recently, Convolutional Neural Networks models were proposed to tackle AU recognition problem [12], [34]. Gudi et al. [12] proposed to use a shallow and small network to recognize the presence and intensity of AUs. Zhao et al. [34] proposed a local patch layer to better model relationships between AUs and certain parts of the face. All of the above work performs direct AU prediction, by using regression or classification on images. We, however, look at differences between expressions to create a ranking of them.

Similar to our work Khademi and Morency [16] proposed to compare the facial appearance in neighboring frames to determine if AU intensity increased, decreased or stayed the same. However, they only performed local comparison of expressions rather than performing global ranking without which absolute AU intensities cannot be reconstructed. Furthermore, they only considered presence and absence of AUs rather than their actual intensities – the focus of our work.

Another area relevant to our work is the use of Preference Learning [10] for predicting affect [19]. This work relies on the ordinal nature of dimensional affect labels (e.g. how happy or excited someone is) to avoid labeler bias during training. However, it is still a direct prediction model that relies on local ranking only during the model training and not testing/inference.

In addition to AU intensity estimation, our work is also related to other approaches that use ranking for model training or during inference. This is common in information retrieval [5], where models would be trained to retrieve a

ranked list relevant to a query. Similar approaches have also been used for recommender systems [17], to refine machine translation results [27], and to perform image captioning [14]. Our work is different from those mentioned above as it does not rely on an original user query to provide a ranking, but instead ranks all of the images in a collection.

3. Local-Global Ranking method

Our Local-Global Ranking method was inspired by league table ranking algorithms and can be seen in terms of competitive games amongst many players [13]. Local comparisons can be seen as individual matches and the global comparison as a tournament. A player (facial expression image) with a higher skill level (AU intensity) is more likely to win in a match against one at a lower skill level, however a loss or a draw are not impossible, just unlikely. After a number of matches between players (a tournament) we can start inferring the true skill of each player by only observing their wins, losses and draws [13]. Our Local-Global Ranking approach uses this intuition; however, instead of players we have images of facial expressions and instead of matches we have a local comparison module. We first describe our approach for creating a local ranking between pairs of expression images (individual matches) and how such local ranks can be integrated using global ranking to predict AU intensities (overall tournament).

We formulate facial action unit intensity recognition as a regression problem, with the goal of learning a function $\mathbf{y} = h(\mathbf{X})$, where $\mathbf{X} = \{\mathbf{x}_i \in \mathcal{R}^d\}$ is a collection of face images $\mathbf{x}_i$ and $\mathbf{y} = \{y_i \in [0, 5]\}$ are the corresponding intensities of an action unit.

3.1. Local ranking

Given a set of n face images from person k with their corresponding labels $\{(\mathbf{x}_i^k \in \mathcal{R}^d); (y_i^k \in [0, 5])\}^n$, we construct a new comparison dataset $T_k = \{(\mathbf{x}_i^k, \mathbf{x}_j^k); (r_{i,j}^k \in \{-1, 0, 1\})\}^m$, of size m. We omit k from further description for brevity. We define $r_{i,j}$ as:

$$r_{i,j} = \begin{cases} 1, & \text{if } y_i > y_j \\ 0, & \text{if } y_i = y_j \\ -1, & \text{if } y_i < y_j \end{cases} \tag{1}$$

where $r_{i,j} = \{-1, 0, 1\}$ represents one of the three comparison outcomes: higher (1), lower (-1), and equal (0). Modeling $r_{i,j}$, can be seen as a three-way classification problem between two observations. Framing the problem in such a way allows us to learn a classifier that is able to compare images rather than just predict the intensity of expression in a single image – an easier task.

We can use a set of person specific datasets T_k to form a full dataset T that can be used to train any three-way classifier that learns to predict $r_{i,j}$ given $\mathbf{x}_i, \mathbf{x}_j$. In our experiments we use a Support Vector Machine (see Section 4.1), but the approach can generalize to other multi-class classifiers as well.

Another advantage of our formulation is that it does not take the size of the difference between labels $(y_i - y_j)$ into account and only cares about their order. This allows it to be robust to the fact that distance in appearance between AU intensity level 1 and 2 is not the same as distance between levels 2 and 3 [24]. This is also the case with a number of other human-produced intensity labels, such as affect [19] and medical symptoms [26]. Our formulation allows us to avoid the non-linear scale bias that assumes that there is a fixed difference between intensity levels of y_i.

3.1.1. Pair-wise sampling. We now describe image sampling techniques to construct the comparison training set T and a set o local pair-wise comparisons $\mathbf{r} = \{r_{i,j}, r_{j,k}, r_{k,l} \ldots\}$ used for global ranking (Section 3.2).

One possibility is to sample each pair of images in a collection, however, this would be computationally infeasible for long video sequences as the number of comparisons would become too large. Instead, given a set of videos from N subjects with F_i number of frames each, we perform the following sampling:

$$T = \bigcup_{n=1}^{N} T_n \tag{2}$$

$$T_n = T_{\text{neigh}} \cup T_{\text{u}} \tag{3}$$

Which is made up of neighborhood sampling (suitable for video sequences):

$$T_{\text{neigh}} = \bigcup_{i=1}^{F_i} \bigcup_{|j-i|<k} \{(\mathbf{x}_i, \mathbf{x}_j); (r_{i,j})\} \tag{4}$$

And uniform sampling (suitable for both video sequences and image collections):

$$T_{\text{u}} = \bigcup_{i=1}^{F_i} \bigcup_{j=\mathcal{U}_d\{1,F_i\}} \{(\mathbf{x}_i, \mathbf{x}_j); (r_{i,j})\} \tag{5}$$

where $\bigcup$ is set union, and $\mathcal{U}_d\{1, F_i\}$ represents d samples from a discrete uniform distribution ranging from 1 to F_i. Such a formulation allows us to sample both distant and nearby images in a video sequence.

We can use T directly as a training set to train a three-way classifier – our local ranking model.

To construct a sequence of observations $\mathbf{r}$ for inference, we perform same sampling as above, but we only sample from one video or collection ($N = 1$). Furthermore, instead of using ground-truth values we use the inferred values of $r_{i,j}$ using our local ranking model.

3.2. Global ranking

The input to our global ranking model is a set of local comparisons $r_{i,j}$ (Section 3.1) between two face images i and j, and the output is the actual AU intensity labels y_i for each image.

In our global ranking we assume that each image i has a true AU intensity $\tilde{y}_i$ associated with it (as ranking does not

have units we indicate the unnormalized intensity using $\sim$). During local comparison it can exhibit observed intensity $o_i = \mathcal{N}(o_i; \tilde{y}_i, \sigma_i^2)$. That is an expression image i exhibits an observed intensity o_i centered around its true intensity $\tilde{y}_i$, with a variance σ_i^2.

We can model the probability of the true intensity given the outcome of their comparison as follows:

$$p(\tilde{y}_i, \tilde{y}_j | r_{i,j}) = \frac{p(r_{i,j} | \tilde{y}_i, \tilde{y}_j) p(\tilde{y}_i, \tilde{y}_j)}{p(r_{i,j})} \tag{6}$$

The prior distribution is modeled as follows: $p(\tilde{y}_i, \tilde{y}_j) = \mathcal{N}(o_i; \tilde{y}_i, \sigma_i^2) \cdot \mathcal{N}(o_j; \tilde{y}_j, \sigma_j^2)$. The probability of a local comparison outcome $r_{i,j}$ between i and j is modeled as:

$$p(r_{i,j} | \tilde{y}_i, \tilde{y}_j) = p(o_i > o_j + \epsilon) \tag{7}$$

The probability $p(r_{i,j} | \tilde{y}_i, \tilde{y}_j)$, is the probability of a comparison outcome, given the true intensities $\tilde{y}_i, \tilde{y}_j$ and their observed intensities o_i, o_j. Including ϵ allows to handle equal expressions (draws) with $|o_i - o_j| < \epsilon$.

Instead of ever actually observing o_i we observe local comparison outcomes $r_{i,j}$ that allow us to infer $\tilde{y}_i$. The true intensities of AU $\tilde{y}_i$ are updated each time an outcome of a local comparison $r_{i,j}$ is observed. This is done using an online learning scheme referred to as Gaussian density filtering that uses message passing over factor graph models [13]. After a set of local comparisons $\mathbf{r} = \{r_{i,j}, r_{j,k}, r_{k,l} \ldots\}$ between expressions we approach their true intensity.

The true expression intensity $\tilde{y}_i$ for each expression i produced by our Local-Global model is unitless and will depend on the initialization of true intensities and their variance (instead it can be seen as a rank rather than actual intensity). To convert it to an actual normalized AU intensity, we use a re-normalization function:

$$y_i = \phi(\tilde{y}_i; \theta) \tag{8}$$

We experimentally found that a linear ϕ transformation from ranks to AU intensities to be sufficient.

Finally, while our global ranking model operates on $y \in [0, 5]$ it generalizes to a broader case where $y \in \mathcal{R}$ or is ordinal.

4. Experiments

We evaluated our Local-Global Ranking method on the task of spontaneous facial action unit intensity estimation in videos. The purpose of the experiments was to compare our model to conventional AU intensity prediction models and modern approaches for within and across dataset AU prediction. In this section we describe the datasets, baselines, and methodology used in our experiments.

We perform three sets of experiments: 1) comparing Local-Global Ranking method to a set of recent baselines; 2) exploring the effect of the pair-wise comparison selection; 3) investigating the cross-dataset generalization of the Local-Global Ranking.

Datasets We performed our experiments on two publicly available datasets for spontaneous facial action unit intensity recognition: Denver Intensity of Spontaneous Facial Action – **DISFA** [20] and Binghamton-Pittsburgh 4D Spontaneous Expression Database – **BP4D** [33]. See Figure 3 for example images from the datasets.

(a) DISFA dataset (b) BP4D dataset

Figure 3: Example frames from the two publicly available datasets used during our evaluations.

The BP4D dataset includes videos of 41 participants – 21 in the official training set, and 20 in the validation set. The dataset is gender balanced and includes annotations for 11 AUs for occurrence and 5 AUs for intensities. In total, the dataset consists of $150,000$ high resolution AU labeled images of spontaneous facial expressions.

DISFA contains videos of 27 participants. It includes 4 minute-long videos of spontaneous facial expression and contains over $130,000$ frames. For every video frame, the intensity of 12 AUs was manually annotated. AUs 6, 12, and 17 have intensity labels across both of the datasets, allowing for cross-dataset evaluation.

Baseline Models We compare our work to a number of recent approaches for action unit intensity estimation.

Valstar et al. [30] proposed the use of an **SVM** model as a baseline for the Facial Expression Recognition (FERA) challenge. Their proposed model uses geometry and Local Binary Gabor Patterns [32] appearance features alongside an SVM.

A similar **HOG-SVR** model was proposed by Baltrušaitis et al. [2]. It uses person-normalized Histogram of Oriented Gradients features alongside geometry ones. As a regressor it employs a linear kernel Support Vector Regressor (SVR).

Continuous Conditional Neural Fields (**CCNF**) model is a temporal approach for AU intensity estimation [1] based on non-negative matrix factorization features around facial landmark points. It performs a non-linear mapping from features to AU predictions with temporal smoothing.

Iterative Regularized Kernel Regression **IRKR** [22] is a recently proposed kernel learning method for AU intensity estimation. It is an iterative nonlinear feature selection method with a Lasso-regularized version of Metric Regularized Kernel Regression. It has shown superior performance over a number of other approaches on the DISFA dataset.

Wang et al. [31] proposed a Bayesian Network (**BN**) that captures the relationship between facial expressions and

AUs. The model relies on the presence of expression labels during model training.

A generative latent tree (**LT**) model was proposed by Kaltwang et al. [15]. The model demonstrates good performance under noisy input.

Finally, we included two recent Convolutional Neural Network (**CNN**) baselines. The shallow four-layer model proposed by Gudi et al. [12], and a deeper CNN model used by Zhao et al. [34] (called ConvNet in their work). The **CNN** model proposed by Gudi et al. [12], consists of three convolutional layers, one fully-connected layer and a final linear layer. The Zhao et al. **D-CNN** model uses five convolutional layers followed by two fully-connected layers of 4096 and 2048 neurons each and a final linear layer.

We report the results of the following models directly from their papers: **SVM**, **CCNF**, **IRKR**, **BN**, and **LT**. We re-implement the other baselines and describe the implementation details in the following section.

Implementation details To extract a face representation we used OpenFace [3] to detect facial landmarks in both of the datasets. We used the landmarks to similarity align all of the face images to a common reference frame and to remove the background in the image.

To train the HOG-SVR model we used a linear kernel Support Vector Regressor (SVR). As input we used concatenated geometry and Histogram of Oriented Gradients (HOG) [9] appearance features extracted using OpenFace. For training we balanced the neutral and expressive frames. As input we used per-person normalized features by subtracting the median feature. For SVR we validated the C parameter in the range $\{10^{-9}, 10^{-8}, \ldots, 10^{-3}\}$.

For both CNN model training we used the similarity aligned images, 48×48 and 170×170 pixels wide for **CNN** and **D-CNN** models respectively. As a loss function, we used Euclidean distance loss. The models were trained using the AdaDelta gradient descent with an initial learning rate of 1.0. For both model training we used dropout [28] on the fully connected layer at the rate of 0.5 for D-CNN, and rate of 0.2 for CNN. For training we balanced the neutral and expressive frames. Early stopping regularization on the validation set was used to avoid over-fitting during the training. We follow the mini-batch training mode, batches of which are set to 100 samples. We use a per-person image normalization by subtracting the mean person image. A separate model was trained for each AU.

4.1. Local-Global Ranking implementation

As a local comparison module we used a linear kernel Support Vector Machine (SVM):

$$r_{i,j} = \mathbf{SVM}(F(\mathbf{x}_i) - F(\mathbf{x}_j)) \tag{9}$$

where $\mathrm{SVM}(\mathbf{x})$ is the SVM prediction function [6], and $F(\mathbf{x})$ are the concatenated appearance and geometry features corresponding to face image $\mathbf{x}$. We used the same features as in the HOG-SVR baseline.

To construct the training data T: for DISFA we set $d = 1$, and $k = 2$; for BP4D we set $d = 2$, and $k = 3$ (see Equations 4 and 5 in Section 3.1.1).

For training the SVM local comparison classifier we balanced positive, negative, and equal comparison samples. We validated the C hyper-parameter in the range $\{10^{-9}, 10^{-8}, \ldots, 10^{-3}\}$. To choose the hyper-parameter we optimized for the mean F1 score of the three-way one vs. all classifier. We trained two separate local-classifiers for comparing neighboring and distant images. Separate local models were trained for each AU.

To construct a set of observations $\mathbf{r}$ that will be used by the global ranker we set $d = 30$, $k = 5$ (see Equations 4 and 5 in Section 3.1.1), this led to 35 pair-wise comparisons per every frame in a video. We also explore the importance of these parameters in one of our experiments (Section 5.2). The true intensities t_i were initialized to 0 and their corresponding variance $\sigma_i^2 = 5$ at the start of Bayesian ranking.

We used a linear function as ϕ in Equation 8. To learn the slope and bias term we first perform Local-Global ranking on the validation set. The resulting ranks t_i are clipped so that $t_i \in [\tau_{\text{low}}, \tau_{\text{high}}]$ ($\tau_{\text{low}}, \tau_{\text{high}}$ are determined by validation). This allows us to train a mapping that is more robust to outliers that can happen due to a frame winning or losing due to problems in tracking/occlusion/pose variation etc. During testing, after applying ϕ we also clip the resulting values y_i to a range of $[0, 5]$ to eliminate outliers.

4.2. Experimental methodology

Within dataset experiments For DISFA experiments we performed person independent 5-fold testing with stratified hold-out validation. Stratified validation reduced the effect of data imbalance by making sure that training and validation sets contain similar numbers of expressive images. For BP4D experiments we used all of the 20 people from the official development set for testing, and the 21 person from the training set was split into 15 training and 6 development subjects (using stratified hold-out validation). As BP4D contains 8 video segments for each of the participants, we concatenated the 8 segments to one super-segment in our experiments (for training/validation and testing).

Cross-dataset experiments We trained AU intensity estimators on BP4D training dataset and tested them on DISFA. The reason we chose to train on BP4D and test on DISFA is because the former contains a more balanced set of AU labels with higher inter-coder agreement [30]. We used the same methodology as above to train the BP4D models, using 21 subjects for training and validation.

The first cross-dataset experiment was on complete cross-dataset generalization by using the BP4D trained models directly and evaluating them on all the DISFA subjects.

The second experiment used some training data (4 participants) from DISFA to adapt the predictions using a simple linear transformation followed by thresholding of the output to be in the range of $[0, 5]$. This was done because different AU intensity datasets have been labeled by different observers that might have slightly different definitions.

TABLE 1: Comparing our model to baselines on the DISFA dataset, results reported as Pearson Correlation Coefficient. Note how our ranking approach outperforms the other baselines, especially for rare AUs - 9, 15, 17, 20. (1) used a different fold split. (2) used 9-fold testing. (3) used leave-one-person-out testing.

Method	AU1	AU2	AU4	AU5	AU6	AU9	AU12	AU15	AU17	AU20	AU25	AU26	Mean
IRKR [22](1)	0.70	0.68	0.68	0.49	0.65	0.43	0.83	0.34	0.35	0.21	0.86	0.62	0.57
LT [15](2)	0.41	0.44	0.50	0.29	0.55	0.32	0.76	0.11	0.31	0.16	0.82	0.49	0.43
CNN [12]	0.60	0.53	0.64	0.38	0.55	0.59	0.85	0.22	0.37	0.15	0.88	0.60	0.53
D-CNN [34]	0.49	0.39	0.62	0.44	0.53	0.55	0.85	0.25	0.41	0.19	0.87	0.59	0.51
CCNF [1](3)	0.48	0.50	0.52	0.48	0.45	0.36	0.70	0.41	0.39	0.11	0.89	0.57	0.49
SVR-HOG [2]	0.64	0.50	0.70	0.67	0.59	0.54	0.85	0.39	0.49	0.22	0.85	0.67	0.59
Local-Global	0.68	0.60	0.78	0.68	0.58	0.64	0.86	0.42	0.52	0.32	0.87	0.65	**0.63**

TABLE 2: Results on BP4D dataset using intra-class correlation (ICC) and Pearson correlation coefficient (PCC).

	PCC						ICC					
Method	AU6	AU10	AU12	AU14	AU17	Mean	AU6	AU10	AU12	AU14	AU17	Mean
SVM-geom. [30]	0.70	0.72	0.71	0.47	0.37	0.59	0.69	0.70	0.65	0.45	0.28	0.55
SVM-app [30]	0.72	0.68	0.70	0.40	0.30	0.56	0.69	0.64	0.67	0.32	0.19	0.50
BN [31]	0.78	0.67	0.84	0.29	0.58	**0.63**	0.77	0.66	0.84	0.20	0.53	0.60
CNN [12]	0.71	0.69	0.84	0.38	0.47	0.62	0.65	0.65	0.84	0.35	0.46	0.59
D-CNN [34]	0.72	0.70	0.77	0.37	0.46	0.60	0.72	0.68	0.76	0.32	0.41	0.58
SVR-HOG [2]	0.73	0.61	0.73	0.40	0.52	0.60	0.71	0.58	0.71	0.33	0.46	0.56
Local-Global	0.76	0.62	0.78	0.43	0.58	**0.63**	0.76	0.62	0.78	0.40	0.55	**0.62**

TABLE 3: Results on BP4D dataset using different pair-wise comparison samples for global ranking construction.

	PCC						ICC					
Method	AU6	AU10	AU12	AU14	AU17	Mean	AU6	AU10	AU12	AU14	AU17	Mean
Neighbors	0.14	0.13	0.14	0.09	0.18	0.14	-0.03	0.08	0.05	0.06	0.07	0.04
Distant	0.75	0.62	0.78	0.43	0.56	**0.63**	0.75	0.61	0.78	0.38	0.53	0.61
Neigh + Distant	0.76	0.62	0.78	0.43	0.58	**0.63**	0.76	0.62	0.78	0.40	0.55	**0.62**

This experiment allows us to test how well our framework would generalize on other datasets if it had access to a small amount of adaptation data. Same adaptation was performed on all of the baselines.

Evaluation For performance evaluation we used commonly accepted metrics in the field [30]: Intra-class Correlation (ICC), Pearson Correlation Coefficient (PCC), and Concordance Correlation Coefficient (CCC) – as different works use different metrics we use the metrics reported by the baselines. PCC assesses how well the predictions follow the ground truth trend, but ignores the shift or scale of the predictions; CCC, and ICC tackle that by including the actual disagreement between labels as well. To compute the evaluation metrics we concatenate the predictions for each test subject (and fold) into a joint vector.

5. Results and discussion

In this section we present the results of our three experiments and demonstrate that our model outperforms state-of-the-art baselines for AU intensity estimation both within (Section 5.1) and across (Section 5.3) datasets. Furthermore, we demonstrate the effect of selection of pair-wise local comparisons in Section 5.2.

5.1. Comparison to baselines

We tested how well our approach compares to state-of-the-art baselines. The results of these comparisons can be seen in Table 1 and Table 2. We outperform the previous baselines on both of the datasets. The gain is particularly large for rarely occuring AUs (9, 14, 15, 17, and 20). which prove to be particularly difficult for the baselines.

We further compare our Local-Global Ranking approach to the SVR-HOG baseline as it used the same features as we did, but performed a direct prediction. For both the DISFA and BP4D datasets a paired sample t-test revealed a statistically significant difference between the mean PCC values of Local-Global Ranking method and SVR-HOG ($p < 0.05$ in both cases). The same trend was observed for CCC metrics (not reported here due to space limitations).

From these results we can conclude that framing the problem of AU intensity estimation as a Local-Global Ranking one improves AU intensity estimation accuracy. We believe this is because it is an easier task than direct prediction and because it allows us to exploit the availability of a number of expression images of the same person. One could argue that gain in performance is due to use measuring the change in expression (by using $F(\mathbf{x}_i) - F(\mathbf{x}_j)$ in Equation 9). However, the models we compare with (SVM, SVR-HOG, CNN, and D-CNN) used neutral (or most common) expression normalized features for prediction as well.

5.2. Pair-wise sampling analysis

We also explored the importance of neighbor and randomly sampled comparisons when constructing $\mathbf{r}$. We compared what happens if we only sample neighboring frames

TABLE 4: Cross-dataset evaluation of different learning paradigms by training on BP4D and testing on DISFA (all 27 subjects). We compare to other approaches, SVR model [2], and a CNN one [12]. LGR is our Local Global Ranking Method. We also compare to the best performing within dataset model (trained and tested on DISFA) – SVR-HOG model. Note that our Local-Global Ranking method generalizes better across datasets.

	PCC					CCC				
	Across				Within	Across				Within
AU	LGR (ours)	SVR [2]	CNN [12]	D-CNN [34]	SVR [2]	LGR (ours)	SVR [2]	CNN [34]	D-CNN [12]	SVR [2]
6	0.57	0.54	0.62	0.56	0.59	0.41	0.37	0.54	0.54	0.53
12	0.75	0.74	0.69	0.68	0.85	0.56	0.62	0.63	0.63	0.82
17	0.59	0.43	0.22	0.20	0.49	0.48	0.33	0.10	0.07	0.42
Mean	**0.64**	0.57	0.51	0.48	0.64	**0.48**	0.44	0.42	0.41	0.59

TABLE 5: Cross-dataset evaluation of different learning paradigms by training on BP4D and testing on DISFA (23 subjects) with adaptation (on 4 subjects). LGR is our Local Global Ranking Method. We also compare to the best performing within dataset model (trained and tested on DISFA) – SVR-HOG model. Note that Local-Global Ranking method generalizes better across datasets, furthermore, it even outperforms the within-dataset model.

	PCC					CCC				
	Across				Within	Across				Within
AU	LGR (ours)	SVR [2]	CNN [12]	D-CNN [34]	SVR [2]	LGR (ours)	SVR [2]	CNN [12]	D-CNN [34]	SVR [2]
6	0.63	0.57	0.60	0.58	0.59	0.62	0.45	0.54	0.53	0.55
12	0.81	0.77	0.69	0.69	0.87	0.81	0.73	0.56	0.56	0.87
17	0.59	0.47	0.26	0.14	0.48	0.54	0.40	0.15	0.13	0.49
Mean	**0.68**	0.61	0.52	0.40	0.65	**0.66**	0.53	0.42	0.40	0.64

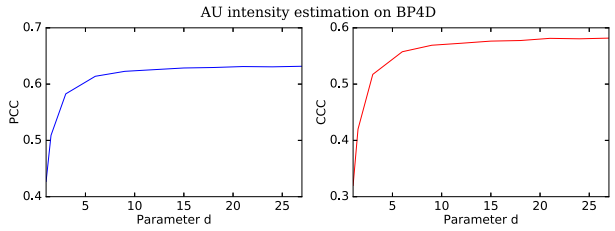

Figure 4: Effect of number of pair-wise comparisons(d) on AU intensity estimation performance on the DISFA dataset. Performance metrics are reported as averages over all 12 AUs. Note how performance increases with extra comparisons, but saturates with as low as 20.

($k = 5, d = 0$), or distant frames ($d = 30, k = 0$) or combination of both ($k = 5, d = 30$) for pair-wise comparison (see Equations 4 and 5 in Section 3.1.1). Results of this experiment can be seen in Table 3. We can conclude two main things from these results. First, sampling both in the neighborhood and in the whole video is beneficial – although the main gain comes from whole video sampling. Second, and more importantly, the strength of our approach does not come simply from local *smoothing* but rather from comparison of images, because a model that just looks at the neighborhood fails. We believe this is because there are few diverse comparison samples for each frame and the expression stays reasonably constant, furthermore, such model is susceptible to drift if some of the comparisons are incorrect.

As a follow up analysis we varied the number of pair-wise uniform comparisons performed (d parameter in Equation 5) when constructing observations for the global ranking. Results of number of comparisons used and the average accuracy of the model across all the AUs on the BP4D dataset can be seen in Figure 4. Note how the model performance saturates (reaches 99% of final accuracy) after 20 pair-wise comparisons – a relatively small number and one that allows for a real-time implementation of Local-Global Ranking.

5.3. Cross-dataset experiments

We tested how well our approach generalizes across datasets. The results of the first cross-dataset experiment can be seen in Table 4. It can be clearly seen that our ranking framework generalizes better across datasets by a significant margin. We believe this is because learning to compare images is more generalizeable than learning how to directly classify them. This is especially true for the PCC metric which looks at the trends rather than the absolute values, indicating that our model is good at capturing them without the need for dataset adaptation.

The results of the adaptation experiment can be seen in Table 5. We can see that ranking again leads to better generalization, we also see that an SVR model is able to generalize better than the CNN ones. More impressively training a ranking model on BP4D and adapting it on DISFA leads to better accuracy than training a regression model on DISFA directly (while the same trend is not observed by adapting other regressors). We believe this is because BP4D is more balanced and has more diverse samples. Our approach is able to exploit that and generalize.

We believe these results are in part because our approach allows us to avoid labeling bias across datasets and individuals. For example, the annotators of one dataset might have slightly different rules about what counts as intensity C ($y = 3$) for an action unit than annotators of another dataset. However, the expressions frames with intensity C will be

more pronounced than those of intensity B ($y = 2$), despite the dataset used. In other words, while absolute values might differ the relative ranks should stay the same.

In conclusion, we demonstrate that learning to rank AU intensities is more generalizeable than learning to predict them directly. We further believe that this approach is able to better handle person and dataset idiosyncrasies.

6. Conclusions

In this paper we presented a novel way of predicting facial action unit intensities using a Local-Global Ranking method. We have demonstrated the utility of our approach on two publicly available datasets. Our approach exhibits significantly better performance than direct AU intensity prediction models. Furthermore, our Local-Global Ranking leads to superior performance for cross-dataset AU prediction when compared to baselines. This is crucial to bring AU recognition to the real world and outside laboratory recorded data. All of these results provide support to our hypothesis that ranking images is an easier task that performing direct prediction – Local-Global Ranking method allows us to exploit that.

Acknowledgments

We acknowledge funding support from the BrainHub Post-doctoral fellowship. The content does not necessarily reect the position or the policy of BrainHub, and no ofcial endorsement should be inferred.

References

[1] T. Baltrušaitis, P. Robinson, and L-P. Morency. Continuous Conditional Neural Fields for Structured Regression. In *ECCV*, 2014.

[2] T. Baltrušaitis, M. Mahmoud, and P. Robinson. Cross-dataset learning and person-specific normalisation for automatic Action Unit detection. In *FERA Challenge, in conjunction with FG*, 2015.

[3] T. Baltrušaitis, P. Robinson, and L-P. Morency. OpenFace: an open source facial behavior analysis toolkit. In *IEEE WACV*, 2016.

[4] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan. Real time face detection and facial expression recognition: Development and applications to human computer interaction. In *CVPRW*, 2003.

[5] Z. Cao, T. Qin, T. Liu, M. Tsai, and H. Li. Learning to Rank: From Pairwise Approach to Listwise Approach. In *ICML*, 2007.

[6] Corinna Cortes and Vladimir Vapnik. Support-Vector Networks. *Machine Learning*, 20:273–297, 1995.

[7] F. De la Torre and J. F Cohn. Facial Expression Analysis. In *Guide to Visual Analysis of Humans: Looking at People*. 2011.

[8] P. Ekman and W. V. Friesen. *Manual for the Facial Action Coding System*. Palo Alto: Consulting Psychologists Press, 1977.

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminative Trained Part Based Models. *TPAMI*, 2010.

[10] Johannes Furnkranz and Eyke Hullermeier. Pairwise Preference Learning and Ranking. In *ECML*, 2003.

[11] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. Mavadati, and D. P. Rosenwald. Social risk and depression: Evidence from manual and automatic facial expression analysis. In *FG*, 2013.

[12] Amogh Gudi, H Emrah Tasli, Tim M Den Uyl, and Andreas Maroulis. Deep Learning based FACS Action Unit Occurrence and Intensity Estimation. In *FERA Challenge, in conjunction with FG*, 2015.

[13] Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill™: A bayesian skill rating system. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *NIPS*. MIT Press, 2007.

[14] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

[15] S. Kaltwang, S. Todorovic, and M. Pantic. Latent trees for estimating intensity of facial action units. In *CVPR*, Boston, MA, USA, 2015.

[16] Mahmoud Khademi and Louis Philippe Morency. Relative facial action unit detection. In *WACV*, 2014.

[17] Y. Lv, T. Moon, P. Kolari, Z. Zheng, X. Wang, and Y. Chang. Learning to Model Relatedness for News Recommendation. In *International Conference on World Wide Web*, 2011.

[18] B. Martinez and M.F. Valstar. Advances, challenges, and opportunities in automatic facial expression recognition. In B. Smolka M. Kawulok, E. Celebi, editor, *Advances in Face Detection and Facial Image Analysis*, pages 63 – 100. Springer, 2016.

[19] Hector P. Martinez, Georgios N. Yannakakis, and John Hallam. Don't Classify Ratings of Affect; Rank Them! *TAC*, 2014.

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

[21] D. McDuff, R. el Kaliouby, D. Demirdjian, and R. Picard. Predicting online media effectiveness based on smile responses gathered over the internet. In *FG*, 2013.

[22] J. Nicolle, K. Bailly, and M. Chetouani. Real-time facial action unit intensity prediction with regularized metric learning. *IVC*, 2016.

[23] P. Robinson and R. el Kaliouby. Computation of emotions in man and machines. *Phil. Trans. of the R. Soc. B: Biological Sciences*, 2009.

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

[25] E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation and recognition. *TPAMI*, 2014.

[26] A. Sarkar, C. Morrison, J. F. Dorn, R. Bedi, S. Steinheimer, J. Boisvert, J. Burggraaff, M. D'Souza, P. Kontschieder, S. Rota Bulò, L. Walsh, C. P. Kamm, Y. Zaykov, A. Sellen, and S. Lindley. Setwise Comparison: Consistent, Scalable, Continuum Labels for Computer Vision. In *CHI*, 2016.

[27] Libin Shen, B C Va, and Marina Rey. Discriminative Reranking for Machine Translation. In *NAACL*, 2004.

[28] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout : A Simple Way to Prevent Neural Networks from Overfitting. *JMLR*, 2014.

[29] T. Tron, A. Peled, and A. Grinsphoon. Automated Facial Expressions Analysis in Schizophrenia : a Continuous Dynamic Approach. In *Pervasive Computing Paradigms for Mental Health*, 2015.

[30] M. Valstar, J. Girard, T. Almaev, G. McKeown, M. Mehu, L. Yin, M. Pantic, and J. Cohn. FERA 2015 - Second Facial Expression Recognition and Analysis Challenge. In *IEEE FG*, 2015.

[31] S. Wang, Q. Gan, and Q. Ji. Expression-assisted facial action unit recognition under incomplete AU annotation. *Pattern Recognition*, 2017.

[32] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang. Local gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face representation and recognition. In *ICCV*. IEEE Computer Society, 2005.

[33] 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.

[34] K. Zhao, W. Chu, and H. Zhang. Deep Region and Multi-Label Learning for Facial Action Unit Detection. In *CVPR*, 2016.