Pose-independent Facial Action Unit Intensity Regression Based on Multi-task Deep Transfer Learning

Yuqian ZHOU, Jimin PI, and Bertram E. SHI
Department of Electronic and Computer Engineering
The Hong Kong University of Science and Technology
Hong Kong SAR

***Abstract*— Facial expression recognition plays an increasingly important role in human behavior analysis and human computer interaction. Facial action units (AUs) coded by the Facial Action Coding System (FACS) provide rich cues for the interpretation of facial expressions. Much past work on AU analysis used only frontal view images, but natural images contain a much wider variety of poses. The FG 2017 Facial Expression Recognition and Analysis challenge (FERA 2017) requires participants to estimate the AU occurrence and intensity under nine different pose angles. This paper proposes a multi-task deep network addressing the AU intensity estimation sub-challenge of FERA 2017. The network performs the tasks of pose estimation and pose-dependent AU intensity estimation simultaneously. It merges the pose-dependent AU intensity estimates into a single estimate using the estimated pose. The two tasks share transferred bottom layers of a deep convolutional neural network (CNN) pre-trained on ImageNet. Our model outperforms the baseline results, and achieves a balanced performance among nine pose angles for most AUs.**

I. INTRODUCTION

Automated facial expression analysis provides important cues for inferring the human intent, which can facilitate human-computer interaction. To better define categories of facial expressions, Paul Ekman et al. proposed a Facial Action Coding System (FACS) [1], which encodes localized facial muscle movements as action units (AUs). AU analysis, including AU occurrence detection and intensity estimation, helps to better evaluate and describe human mental states, like depression [2] and happiness [3]. Previous research has mostly studied datasets with near-frontal poses and posed expressions, such as CK+ [4], MMI[5] etc. In reality, non-frontal views of spontaneous facial expressions are common, and present challenges for accurate AU estimation. The FG 2017 Facial Expression Recognition and Analysis challenge (FERA 2017) [6] seeks to address the problem of AU occurrence and intensity estimation on data derived from multi-view spontaneous facial expression videos in the BP4D dataset [7, 8].

In this paper, we focus on the AU intensity estimation sub-challenge. We propose two transfer learning models based on VGG16 [9], a deep neural network trained on a subset of the ImageNet database. The first model is trains a single pose-invariant recognizer, which cascades VGG16 with a randomly initialized two-layer fully connected regression network. The weights of the VGG16 network/two-layer regressor are fine-tuned/trained using images with all different pose angles. The second model trains three pose-dependent regressors and one pose estimator. Both the regressors and the pose estimator share the same bottom layers, which are initialized using the VGG16 weights, but then fine-tuned during training. Our experimental results for both networks outperform the baseline results. In most cases, the multi-task network merging pose-dependent estimates performs better than the single-task pose-invariant network, even without data augmentation.

This work was supported in part by the General Research Fund of the Hong Kong Research Grants Council, under grant 618713.

II. RELATED WORK

AU intensity estimation is an important part in facial expression recognition. McKeown et al. [10] stated that the intensity of facial expression, which is an important feature to distinguish real high-level emotional states from pretended low-level social behavior, helps to assess human psychological states. Common approaches to estimate intensity can be categorized into classification-based and regression-based methods, which can both be further subdivided into static or dynamic models.

Previous efforts for AU intensity estimation using hand-crafted features often use a similar pipeline. Features extracted can be geometric features computed from tracked facial landmarks, or appearance features filtered by local descriptors like LBP [11, 12], Gabor [13] etc. Selected features are further fed into a SVC [14] or AdaBoost [13] for classification, or into a SVR [15-17] or a neural network [15] for regression. These can be extended to dynamic models using the HMM [18], Dynamic Bayesian Network (DBN) [19] or Conditional Random Fields (CRF) [6, 20] etc. These approaches have been evaluated on public benchmark databases like CK-Enhanced [21], DISFA [22], UNBC-McMaster [23], BP4D [7, 8] or SEMAINE [24] etc.

Deep convolutional neural network (CNN) have had great success in many computer vision tasks. Structures like AlexNet [25] and VGG [9] have been proven to have good performance in object recognition. Many past works have applied CNNs to AU detection and intensity analysis. A smile detector was proposed in [3] based on CNN. This was further extended to a dynamic model using a recurrent neural network (RNN). In [26], a single multi-label deep CNN was trained to estimate and classify five AUs simultaneously. Jaiswal et al. [27], who achieved the best performance on the FERA 2015 dataset [12], proposed a single-label network for each AU, and learned the shape, appearance and dynamic features jointly using a deep CNN and bidirectional long-short term memory (Bi-LSTM). However, training a deep CNN requires large-scale data and a

978-1-5090-4023-0/17 $31.00 © 2017 IEEE
DOI 10.1109/FG.2017.112

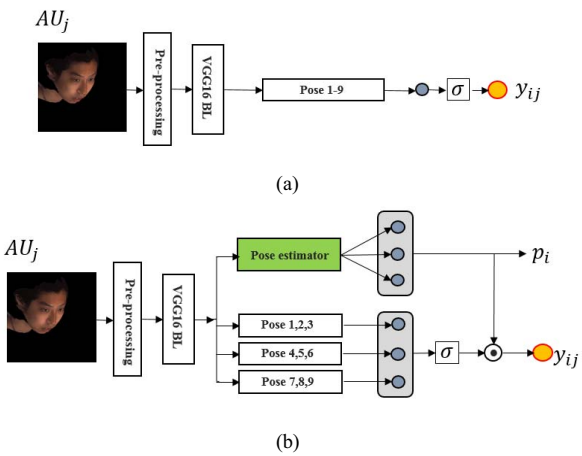


Fig.1. The architecture of the two proposed models. (a) The pose-invariant model concatenates a pre-trained VGG16 network with a randomly initialized two-layer regressor whose output is further rectified by a sigmoid function. The entire network is re-trained for each AU independently using data from all poses according to the Euclidean loss. (b) The pose-dependent model trains three pose-dependent regressors and an auxiliary pose estimator. The nine poses are separated into three groups for training the three pose-dependent regressors. The output of the pose estimator is a winner-take-all network with three units, one for each pose group. The estimator and regressors share the same bottom layer. The three pose estimates for each AU are merged by taking their dot product with the pose estimator output.

long training time. With many parameters and insufficient data, such networks may easily over-fit subjects or specific AUs.

Due to the large demand for data in training deep neural networks, transfer learning has been widely investigated to facilitate the training process with less data. Yosinski et al. [28] systematically evaluated the generality or specificity of deep features, and quantified their transferability. They demonstrated that networks initialized with transferred weights and fine-tuned on new tasks displayed better generalization than networks trained from random weight initialization. Networks that re-use and select deep features pre-trained on non-face large-scale images have outperformed networks trained from scratch in facial expression recognition [29].

Deep transferred CNNs have shown the ability to learn pose-specific representations of faces. In [30], Abd-Almageed et al. applied multiple transferred CNNs trained on rendered faces with different head poses to represent faces, and concatenated the features for face recognition. In the area of AU analysis, Tősér et al. [31] trained a deep CNN to detect AUs on frontal-view faces and evaluated it on the multi-pose augmentation dataset of BP4D. They claimed that the detection performance using the CNN did not degrade significantly with varying head pose. This inspired us to use deep transfer networks to learn pose-specific features for AU intensity estimation.

III. METHODOLOGY

In this section, we introduce our proposed model to address multi-pose AU intensity estimation. First, we introduce the BP4D database used in the FERA 2017 challenge and the methods we used to sample and pre-process the data. Second, we propose two deep network structures and explain the intuition behind the modeling.

A. Dataset

The dataset used for training and evaluation in the FERA 2017 challenge is derived from the 3D model of the BP4D dataset. It contains a total of 41 subjects (18 men and 23 women) in the training partition, and 20 additional subjects (13 men and 7 women) in the validation partition. Subjects show spontaneous facial expressions while performing eight different tasks. For each video sequence, facial images at nine different poses are generated from the 3D model of BP4D.

For the AU intensity estimation sub-challenge, each frame is labelled with the intensity over seven different AUs (AU1, AU4, AU6, AU10, AU12, AU14, and AU17), ranging from zero to five. Frames with missing labels are assigned a value of nine, and are not used for training nor evaluation. More details of the dataset are shown in the baseline paper [6].

For each AU and each of the nine poses, we randomly sampled 3000 non-zero frames (with AU labels from one to five), and 3000 zero frames (with AU label zero) from the training partition as the training set in the experiment. Thus, AU-specific training set contains 54000 images. Similarly, we also sampled 54000 images for each AU as the validation set, which we used to determine when to stop training. The labels of each AU were rescaled to 0-1 for training.

We pre-processed the images by converting them to greyscale, equalizing the histogram and finally resizing them to 224×224. We did not detect the facial landmarks to align the face, since detection of facial landmarks for extreme poses is not reliable. In addition, the nonlinear warping of the face required for extreme poses may adversely affect the appearance shape. Therefore, we only considered the appearance features learned by deep structures from the original images.

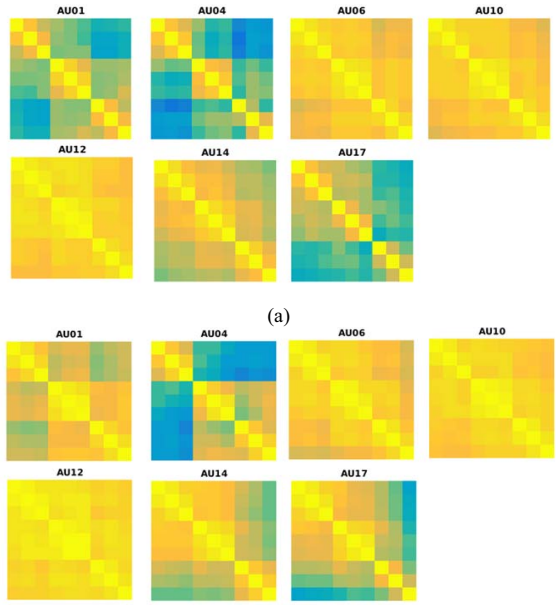
B. Pose-invariant Model

Figure 1(a) shows the architecture of the proposed pose-invariant model we designed in our experiment. In this model, for each AU we concatenate bottom five stages of VGG16 with a randomly initialized two-layer regression network. The VGG16 network takes as input 224×224 pixel images. The bottom five stages include both convolutional and max pooling layers, and output 512 feature maps with size 7×7. The fully connected regression network has 256 neurons in the hidden layer. The output of the regressor is further rectified by a sigmoid function. The entire network is trained using data from one AU but all poses to minimize the Euclidean loss.

Intuitively, given the rich and complex filter banks in the high-level layers of VGG16, the network will be able to learn the most relevant feature maps and capture enough expression variations in terms of facial patterns. By fine-tuning the VGG layers, the network will learn more pose-invariant AU-specific and region-specific features. On the other hand, forcing the network to generate a single estimate across all poses may interfere with its ability to exploit pose-specific features for each AU.

C. Pose-dependent Model

Figure 1(b) illustrates the system architecture of the pose-dependent model. For each AU, instead of training only one regressor for all poses, we separate the nine poses into three



(a)

(b)

Fig.2. The ICC confusion matrix for two models. The color map from blue to yellow corresponds to ICC values from 0 to 1. (a) Pose-invariant model: for AU6, AU10 and AU12, the estimated sequences of each pose are more consistent with each other. The estimation sequences are influenced more by pitch than yaw. (b) Pose-dependent model: consistency is improved for AU1, AU4, AU6, AU10, AU12 and AU17.

groups (pitch up, pitch down and without pitch rotation), and train three pose-dependent regressors. The reason for the grouping criteria will be explained in section IV. We also train an auxiliary pose estimator, which has three softmax outputs approximating the probability that the input image belongs to each pose group. The three regressors and pose estimator share the same bottom layers. The final AU estimate is the dot product between the output of the pose network after winner-take-all (WTA) and the vector of the three pose-dependent AU intensity estimates.

For all AUs except AU12, we jointly trained the three regressors and fine-tuned the bottom layers using data from all poses and AU intensities under the assumption that the output of the pose estimation network is equal to the known pose. This ensures that each network is trained only on data from the corresponding pose group. For AU12, we additionally trained the randomly initialized upper layers of the pose estimator while leaving the weights of the bottom layers unchanged. Finally, we fine-tune the entire AU12 network using the multi-task objective defined by

$$\underset{W^m, W^a}{argmin} \sum_{i=1}^{N} l^m\left(y_i, f\left(x_i; W_m\right)\right) + \sum_{i=1}^{N} \alpha\, l^a\left(p_i, f\left(x_i; W_a\right)\right) \quad (1)$$

where N is the batch size, x_i is the input image, y_i and p_i are the ground truth of AU intensity and pose, W_m and W_a are the weights for AU and pose estimation, l^m is defined by the Euclidean loss, l^a is defined with multi-class cross-entropy loss, and α is the parameter to balance the importance of tasks. In our experiment, α is set to 0.05. We train the network by Mini-batch Stochastic Gradient Descent (SGD) with learning rate 1e-4 in Tensorflow [32], and stop training when the validation loss starts to increase. Finally, we conduct noise suppression on the raw estimated result by median filtering with window length three.

Intuitively, AU intensity estimation at different poses is based on different visual cues, and separating groups of poses to train the network will make the cues more explicit. On the other hand, the increased number of parameters may lead to over-fitting, especially in the absence of data augmentation.

IV. Experimental Results

In this section, we show the experimental results of two systems on the validation set and compare it with the baseline paper. We also visualize the receptive fields of the pose-invariant system to understand the learned features. Finally, we analyze the tradeoff between these two models.

A. Pose-invariant Model

We trained the pose-invariant model on the sampled training set without using pose prior and evaluated the model on the validation and testing partitions in terms of Intraclass Correlation Coefficient (ICC), Pearson Correlation Coefficient (PCC) and Root Mean Square Error (RMSE). The evaluation results on the validation partition are summarized in Table I and the results per pose are also listed in Table II. Compared to the baseline results, the pose-invariant model performs better on every AU and on all evaluation metrics. The overall ICC is improved by about 220%.

We speculate that this improvement is due primarily to two factors. First, the baseline method detects facial landmarks, which has a high failure rate on extreme poses (up to 33.13% for images of pose 1). In contrast, our system does not miss frames when using appearance features. Second, warping the landmarks and aligning them may change the shape of facial structures. In our model, we do not warp the images but learn both the pose and expression variants using rich filter banks to avoid changing the original data.

Intuitively, errors by the pose-invariant model may be partly due to differences in the spatial locations of important cues for AU intensity in the different poses. To validate this intuition, we performed an error analysis by computing the ICC confusion matrix C, whose elements C_{ij} are the ICC values between pairs of poses (i and j). This matrix measures the consistency between estimated sequences of different poses. Figure 2(a) shows the ICC confusion matrix for each AU. We note several observations. First, for AU6, AU10 and AU12, the estimated sequences from different poses are more consistent with each other. Second, for each AU, the estimation sequences are influenced more by pitch than yaw rotations. This inspired us to group each pitch position (e.g. poses 1, 2 and 3) together and train pose-specific models.

In order to validate the match between the learned features and the definition of an AU, we further visualized the receptive fields of the final activation in a single-label regressor under each AU and pose, using the method of deconvolution proposed by Zeiler et al. [33]. Instead of visualizing the high-level feature maps, we back-propagated the final neuron to the pixel space

TABLE I RESULTS COMPARISON FOR THE THREE MODELS ON THE VALIDATION PARTITION

AU	Baseline Model			Pose-invariant Model			Pose-dependent Model		
	RMSE	PCC	ICC	RMSE	PCC	ICC	RMSE	PCC	ICC
AU01	1.006	0.097	0.082	0.548	0.444	0.444	**0.518**	**0.538**	**0.536**
AU04	1.296	0.084	0.069	**0.562**	0.273	0.251	0.571	**0.415**	**0.412**
AU06	1.648	0.429	0.429	0.993	**0.736**	**0.724**	**0.987**	0.718	0.704
AU10	1.628	0.435	0.434	**0.877**	**0.794**	0.774	0.943	0.783	**0.779**
AU12	1.345	0.543	0.540	0.908	0.802	0.800	**0.825**	**0.827**	**0.816**
AU14	1.637	0.264	0.259	**1.140**	**0.607**	**0.548**	1.170	0.591	0.504
AU17	1.256	0.052	0.005	0.865	0.345	0.337	**0.746**	**0.465**	**0.454**
Mean	1.402	0.265	0.260	0.842	0.571	0.554	**0.823**	**0.620**	**0.601**

TABLE II POSE-INVARIANT MODEL RESULTS ON VALIDATION PARTITION, PER POSE

RMSE										
Pose	1	2	3	4	5	6	7	8	9	VI
AU01	**0.533**	0.582	0.578	**0.461**	0.464	0.508	0.594	0.571	0.618	0.548
AU04	**0.628**	**0.625**	**0.640**	0.556	**0.495**	**0.540**	0.547	0.489	0.513	**0.562**
AU06	**1.064**	**1.000**	**1.007**	**0.901**	**0.876**	**0.902**	**1.074**	0.995	1.091	0.993
AU10	0.877	**0.892**	**0.926**	**0.845**	**0.840**	**0.843**	**0.912**	**0.857**	**0.897**	**0.877**
AU12	0.910	0.942	0.991	0.817	0.830	0.851	0.976	0.899	0.940	0.908
AU14	1.166	1.146	1.161	1.125	**1.090**	**1.056**	**1.164**	**1.151**	**1.193**	**1.140**
AU17	0.937	0.973	0.990	**0.744**	0.772	0.848	0.758	0.862	0.858	0.865
mean	0.874	0.880	0.899	**0.778**	**0.767**	**0.793**	0.861	0.832	0.873	0.842
ICC										
Pose	1	2	3	4	5	6	7	8	9	VI
AU01	0.442	0.397	0.434	0.518	0.543	0.546	0.362	0.364	0.424	0.444
AU04	0.183	0.142	0.160	0.321	0.383	0.314	0.235	0.309	0.268	0.251
AU06	**0.722**	**0.744**	**0.722**	**0.767**	**0.770**	**0.744**	0.676	0.709	0.686	**0.724**
AU10	0.768	0.761	0.748	0.789	0.797	0.799	0.760	**0.785**	**0.765**	0.774
AU12	0.798	0.787	0.752	0.841	0.838	0.821	0.769	0.806	0.787	0.800
AU14	0.541	0.559	0.540	0.552	**0.600**	**0.610**	**0.519**	**0.530**	**0.485**	**0.548**
AU17	0.309	0.326	0.346	0.439	0.403	0.360	0.352	0.302	**0.237**	0.337
mean	0.538	0.531	0.529	0.604	0.619	0.599	0.525	0.544	0.522	0.554

TABLE III POSE-DEPENDENT MODEL RESULTS ON VALIDATION PARTITION, PER POSE

RMSE										
Pose	1	2	3	4	5	6	7	8	9	VI
AU01	0.565	**0.557**	**0.528**	0.482	**0.459**	**0.504**	**0.495**	**0.528**	**0.539**	**0.518**
AU04	0.695	0.677	0.745	**0.499**	0.537	0.577	**0.431**	**0.413**	**0.460**	0.571
AU06	1.068	1.027	1.010	0.982	0.944	0.933	0.997	**0.925**	**0.993**	**0.987**
AU10	**0.839**	0.894	0.969	0.916	0.926	0.945	1.013	1.000	0.977	0.943
AU12	**0.855**	**0.847**	**0.837**	**0.765**	**0.738**	**0.763**	**0.937**	**0.826**	**0.842**	**0.825**
AU14	**1.070**	**1.110**	**1.139**	**1.073**	1.117	1.132	1.235	1.309	1.316	1.170
AU17	**0.712**	**0.695**	**0.671**	0.776	**0.771**	**0.790**	**0.684**	**0.752**	**0.848**	**0.746**
mean	**0.829**	**0.830**	**0.843**	0.785	0.785	0.806	**0.827**	**0.822**	**0.854**	**0.823**
ICC										
Pose	1	2	3	4	5	6	7	8	9	VI
AU01	**0.492**	**0.494**	**0.527**	**0.596**	**0.610**	**0.570**	**0.524**	**0.500**	**0.520**	**0.536**
AU04	**0.230**	**0.213**	**0.224**	**0.533**	**0.524**	**0.498**	**0.525**	**0.603**	**0.551**	**0.412**
AU06	0.663	0.680	0.689	0.705	0.725	0.722	**0.725**	**0.737**	**0.694**	0.704
AU10	**0.804**	**0.781**	**0.751**	**0.800**	**0.800**	**0.784**	**0.785**	0.783	0.763	**0.779**
AU12	**0.800**	**0.803**	**0.808**	**0.843**	**0.853**	**0.840**	**0.786**	**0.827**	**0.817**	**0.816**
AU14	**0.584**	**0.571**	**0.550**	**0.578**	0.545	0.522	0.469	0.388	0.363	0.504
AU17	**0.507**	**0.514**	**0.497**	**0.497**	**0.466**	**0.459**	**0.493**	**0.419**	0.235	**0.454**
mean	**0.583**	**0.579**	**0.578**	**0.650**	**0.646**	**0.628**	**0.615**	**0.608**	**0.563**	**0.601**

by computing the gradients of the final output with respect to the original pixels, which is $D_i = \frac{\partial f(x_i;W_m)}{\partial x_i}$, where x_i is the input image, $f(x_i;W_m)$ is the network function of the deep CNN, and D_i is the reconstructed salient receptive fields. We computed D_i over 1000 random samples with highly activated estimated values for each AU and pose, and then summed and normalized them. The visualization results for some AUs are shown in Figure 3, which demonstrates that the network learned to focus on AU-relevant spatial regions regardless of pose. Interestingly, the receptive fields tended to focus on the left half of the face because this region is common among all images from different poses.

B. *Pose-dependent Model*

We grouped the data into pitch up (poses 1, 2 and 3), no pitch (poses 4, 5 and 6) and pitch down (poses 7, 8 and 9), and trained the pose-dependent model without augmenting the original training data. The affiliated pose estimator was trained on top of the AU12 detector, and achieved a 100% classification accuracy on the validation set after jointly training with the AU

TABLE IV RESULTS COMPARISON ON THE TESTING PARTITION

AU	Baseline Model		Proposed Mixed Model	
	RMSE	ICC	RMSE	ICC
AU01	1.082	0.035	**0.744**	**0.307**
AU04	1.200	-0.004	**0.465**	**0.147**
AU06	1.604	0.461	**0.968**	**0.671**
AU10	1.548	0.451	**1.054**	**0.735**
AU12	1.339	0.518	**0.866**	**0.793**
AU14	1.422	0.037	**1.210**	**0.147**
AU17	1.626	0.020	**0.845**	**0.319**
Mean	1.403	0.217	**0.879**	**0.446**

TABLE V MIXED MODEL RESULTS ON TESTING PARTITION, PER POSE

RMSE										
Pose	1	2	3	4	5	6	7	8	9	Global
AU01	0.605	0.717	0.762	0.671	0.722	0.843	0.674	0.761	0.894	0.744
AU04	0.396	0.449	0.560	0.503	0.518	0.542	0.387	0.384	0.407	0.465
AU06	1.046	0.961	0.967	0.975	0.946	0.938	0.977	0.983	0.910	0.968
AU10	0.949	0.901	0.905	1.102	1.110	1.041	1.178	1.171	1.081	1.054
AU12	0.843	0.841	0.825	0.860	0.801	0.771	1.002	0.910	0.918	0.866
AU14	1.306	1.322	1.309	1.075	1.020	1.125	1.252	1.235	1.204	1.210
AU17	0.780	0.758	0.707	0.828	0.870	0.838	0.936	1.002	0.852	0.845
mean	0.846	0.850	0.862	0.859	0.855	0.871	0.915	0.921	0.895	0.879
ICC										
Pose	1	2	3	4	5	6	7	8	9	Global
AU01	0.431	0.355	0.303	0.334	0.315	0.286	0.268	0.272	0.253	0.307
AU04	0.276	0.194	0.113	0.123	0.140	0.130	0.105	0.135	0.141	0.147
AU06	0.624	0.658	0.635	0.609	0.632	0.625	0.742	0.736	0.752	0.671
AU10	0.731	0.762	0.772	0.728	0.732	0.739	0.736	0.733	0.732	0.735
AU12	0.795	0.799	0.802	0.793	0.815	0.822	0.763	0.788	0.782	0.793
AU14	0.221	0.249	0.256	0.121	0.145	0.075	0.075	0.025	0.031	0.147
AU17	0.340	0.377	0.395	0.347	0.341	0.342	0.260	0.245	0.280	0.319
mean	0.488	0.485	0.468	0.436	0.446	0.431	0.421	0.419	0.424	0.446

estimators. We evaluated the system on all the videos in the validation partition (Table I through Table III).

Compared with the pose-invariant model, the performance for most AUs in terms of ICC improved greatly, especially for AU1, AU4, AU12 and AU17, but it was not improved significantly, and even dropped for AU6, AU10 and AU14. In Figure 2(b), the consistency between each pose is adjusted for AU1 and AU17, but becomes worse for AU14.

For testing, for each AU we chose the best model (either pose invariant or pose variant) based on the ICC computed for the validation partition, resulting in an overall ICC value of 0.610 on the validation partition. The evaluation results on the testing partition computed by this mixed model are listed in Table IV and Table V. The system performance is significantly better than that of the baseline paper.

V. CONCLUSIONS

We proposed two transfer-learning models to address multi-pose AU intensity estimation in the FERA 2017 sub-challenge. Both models used fine-tuned bottom layers initialized using the VGG16 network weights. The first contains a single pose-invariant classifier. The second merges results from pose-dependent classifiers. The pose-dependent model can learn the shared pose-specific features and balance the performance across poses. However, the system performance depends critically on the reliability of the pose estimator. For this dataset, poses were easy to differentiate, but this may not be the case for faces in the wild. Increasing the parameter size prolongs the training time, and often requires data augmentation to avoid over-fitting. Our results demonstrate that a deep network can be transferred to learn AU intensity estimation across multiple poses and can greatly outperform the baseline trained on geometric features.

REFERENCES

[1] P. Ekman and W. V. Friesen, "Facial action coding system," 1977.

[2] J. F. Cohn et al., "Detecting depression from facial actions and vocal prosody," in 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 2009, pp. 1-7.

[3] P. O. Glauner, "Deep convolutional neural networks for smile recognition," *arXiv preprint arXiv:1508.06535,* 2015.

[4] 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 *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 2010, pp. 94-101.

[5] M. Valstar and M. Pantic, "Induced disgust, happiness and surprise: An addition to the MMI facial expression database," in *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, 2010, p. 65.

[6] M. F. Valstar *et al.*, "FERA 2017-addressing head pose in the third Facial Expression Recognition and Analysis Challenge," *arXiv preprint arXiv:1702.04174,* 2017.

[7] X. Zhang *et al.*, "BP4D-Spontaneous: a high-resolution spontaneous 3D dynamic facial expression database," *Image and Vision Computing,* vol. 32, no. 10, pp. 692-706, 10// 2014.

[8] Z. Zhang *et al.*, "Multimodal Spontaneous Emotion Corpus for Human Behavior Analysis," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3438-3446.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556,* 2014.

[10] G. McKeown, I. Sneddon, and W. Curran, "Gender differences in the perceptions of genuine and simulated laughter and amused facial expressions," *Emotion Review,* vol. 7, no. 1, pp. 30-38, 2015.
[11] G. Sandbach, S. Zafeiriou, and M. Pantic, "Markov random field structures for facial action unit intensity estimation," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 738-745.
[12] M. F. Valstar *et al.*, "FERA 2015 - second Facial Expression Recognition and Analysis challenge," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2015, vol. 06, pp. 1-8.
[13] J. Hamm, C. G. Kohler, R. C. Gur, and R. Verma, "Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders," *Journal of Neuroscience Methods,* vol. 200, no. 2, pp. 237-256, 9/15/ 2011.
[14] Z. Hammal and J. F. Cohn, "Automatic detection of pain intensity," presented at the 14th ACM International conference on Multimodal interaction, Santa Monica, California, USA, 2012.
[15] Y. Zhang, L. Zhang, and M. A. Hossain, "Adaptive 3D facial action intensity estimation and emotion recognition," *Expert Systems with Applications,* vol. 42, no. 3, pp. 1446-1464, 2/15/ 2015.
[16] A. Savran, B. Sankur, and M. Taha Bilge, "Regression-based intensity estimation of facial action units," *Image and Vision Computing,* vol. 30, no. 10, pp. 774-784, 10// 2012.
[17] S. Kaltwang, "Regression-based estimation of pain and facial expression intensity," Imperial College London, 2015.
[18] S. M. Mavadati and M. H. Mahoor, "Temporal facial expression modeling for automated action unit intensity measurement," in *2014 22nd International Conference on Pattern Recognition*, 2014, pp. 4648-4653.
[19] K. P. Murphy, "Dynamic Bayesian networks: Representation, inference and learning," University of California, Berkeley, 2002.
[20] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning, ICML*, 2001, vol. 1, pp. 282-289.
[21] L. A. Jeni, J. M. Girard, J. F. Cohn, and F. D. L. Torre, "Continuous AU intensity estimation using localized, sparse facial feature space," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1-7.
[22] 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,* vol. 4, no. 2, pp. 151-160, 2013.
[23] 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 *Face and Gesture 2011*, 2011, pp. 57-64.
[24] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The SEMAINE database: annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing,* vol. 3, no. 1, pp. 5-17, 2012.
[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," presented at the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, Nevada, 2012.
[26] A. Gudi, H. E. Tasli, T. M. d. Uyl, and A. Maroulis, "Deep learning based FACS Action Unit occurrence and intensity estimation," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2015, vol. 06, pp. 1-5.
[27] S. Jaiswal and M. Valstar, "Deep learning the dynamic appearance and shape of facial action units," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1-8.
[28] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," presented at the 27th International Conference on Neural Information Processing Systems, Montreal, Canada, 2014.
[29] Y. Zhou and B. E. Shi, "Action unit selective feature maps in deep networks for facial expression recognition," in *Proceedings of IEEE International Joint Conference on Neural Networks*, 2017.
[30] W. AbdAlmageed *et al.*, "Face recognition using deep multi-pose representations," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1-9.
[31] Z. Tősér, L. A. Jeni, A. Lőrincz, and J. F. Cohn, "Deep learning for facial action unit detection under large head poses," in *Computer Vision – ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III*, G. Hua and H. Jégou, Eds. Cham: Springer International Publishing, 2016, pp. 359-371.
[32] M. Abadi *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467,* 2016.
[33] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 818-833.

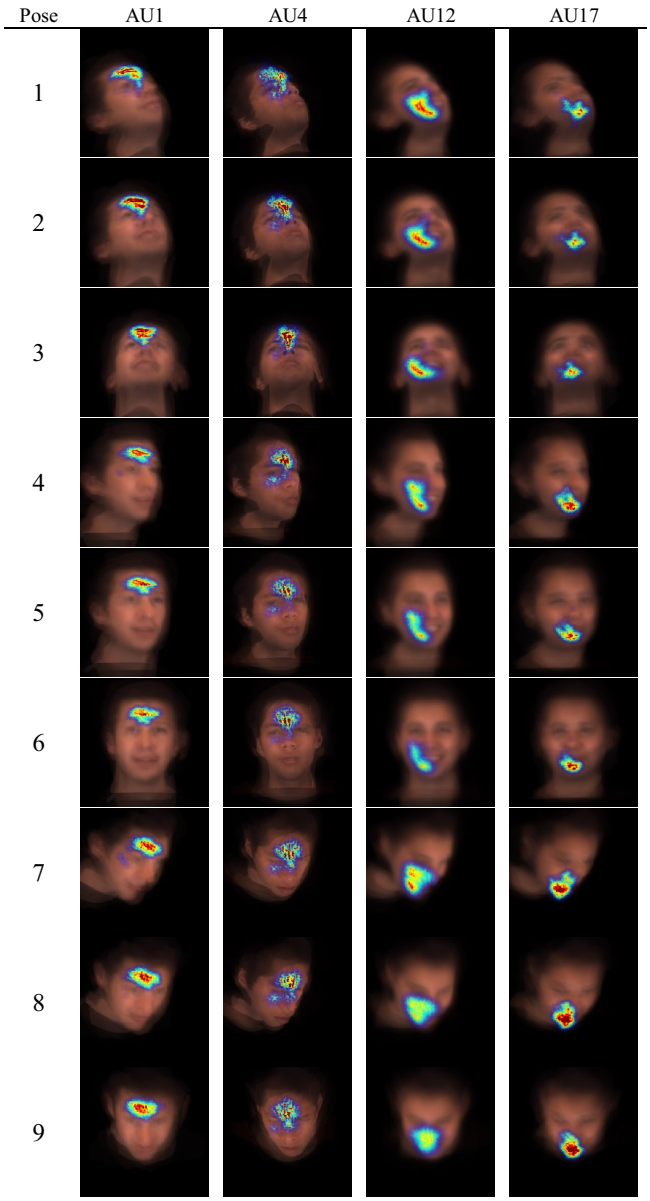


Fig.3. The visualization of learned receptive field of the final regression neuron for some AUs and poses. The network learned to focus on the AU-defined regions regardless of pose.