Recognition of Action Units in the Wild with Deep Nets and a New Global-Local Loss

C. Fabian Benitez-Quiroz   Yan Wang   Aleix M. Martinez
Dept. Electrical and Computer Engineering
The Ohio State University
{benitez-quiroz.1,wang.9021,martinez.158}@osu.edu

1. Derivations of Eq. (9)

The main results given in (9) in the main paper was derived using Green's Theorem. Recall, that Green's Theorem is a mechanism to convert a line integral about a Jordan curve $\mathcal{C}$ to a double integral over the plane $\mathcal{D}$ bounded by $\mathcal{C}$. The Jordan curve $\mathcal{C}$ must be positive oriented and piecewise smooth.

We begin by defining the area of $\mathcal{D}$ we wish to compute. We assume we are in $\mathbb{R}^2$, defined by the (x_1, x_2) coordinate system. Using Green's Theorem, we have

$$\text{Area(D)} = \frac{1}{2}\oint_{\mathcal{C}} -x_2 \mathrm{d}x_1 + x_1 \mathrm{d}x_2, \tag{S1}$$

where $\mathcal{D}$ is a non-self-intersecting polygon bounded by $\mathcal{C}$.

In our case, this polygon and bounding Jordan curve $\mathcal{C}$ are defined by a set of t points on $\mathcal{C}$. An example of this curve was shown in Figure 3 (right-most image) in the main paper. These points were defined as $(\tilde{\mathbf{x}}_{i1}, \dots, \tilde{\mathbf{x}}_{it})$, where $\tilde{\mathbf{x}}_{ik} = (\tilde{\mathbf{x}}_{ik1}, \tilde{\mathbf{x}}_{ik2})^T$.

We call the line segment from each $\tilde{\mathbf{x}}_{ik}$ to $\tilde{\mathbf{x}}_{i(k+1)}$, Γ_i. A parametrization of Γ_i is given by the function $\gamma_i : [0,1] \to \mathbb{R}^2$ and, hence, we have $\gamma_i(a) = (\tilde{x}_{ik1}, \tilde{x}_{ik2}) + a\left(\tilde{x}_{i(k+1)1} - \tilde{x}_{ik1}, \tilde{x}_{(k+1)2} - \tilde{x}_{ik2}\right)$.

We can now compute the line integral of this curve,

$$\begin{aligned}
\oint_{\Gamma_i} -x_2 \mathrm{d}x_1 + x_1 \mathrm{d}x_2 = & \\
& \int_0^1 \left(-\tilde{x}_{ik2} - a\left(\tilde{x}_{(k+1)2} - \tilde{x}_{ik2}\right)\right)\left(\tilde{x}_{i(k+1)1} - \tilde{x}_{ik1}\right)\mathrm{d}a \\
+ & \int_0^1 \left(-\tilde{x}_{ik1} - a\left(\tilde{x}_{(k+1)1} - \tilde{x}_{ik1}\right)\right)\left(\tilde{x}_{i(k+1)2} - \tilde{x}_{ik2}\right)\mathrm{d}a \\
= & \int_0^1 \left(-\tilde{x}_{ik2}\tilde{x}_{i(k+1)1} + \tilde{x}_{ik1}\tilde{x}_{i(k+1)2}\right)\mathrm{d}a \\
= & \; \tilde{x}_{ik1}\tilde{x}_{i(k+1)2} - \tilde{x}_{ik2}\tilde{x}_{i(k+1)1}.
\end{aligned} \tag{S2}$$

Hence, the area of $\mathcal{D}$ is

$$\frac{1}{2}\left[\left(\sum_{k=1}^{t-1}\left(\tilde{x}_{ik1}\tilde{x}_{i(k+1)2} - \tilde{x}_{ik2}\tilde{x}_{i(k+1)1}\right)\right) + \left(\tilde{x}_{it1}\tilde{x}_{i12} - \tilde{x}_{i12}\tilde{x}_{it1}\right)\right]. \tag{S3}$$

2. Extended Experimental Results

The main paper presented comparative results with state-of-the-art algorithms using the *Final score* of the EmotioNet challenge, given in (18). Herein, we provide results with the two metric defining this final score: F_1 score and *Accuracy*, given in (16) and (17), respectively. Figures S1, S2 and S3 plot the *Accuracy* values of the experiments described in Sections 4.1 and 4.2. Figures S4, S5 and S6 show the F_1 scores of these same experiments. Tables S1, S2, S3 and S4 show the results from figures 7-9 in the main manuscript. A right tailed t-test shows that our method is statistically better than state-of-the-art algorithms (p<.0005, p<.005 and p<10^{-8} with respect to JHU, I2R-CCNU-NTU-2 and AlexNet).

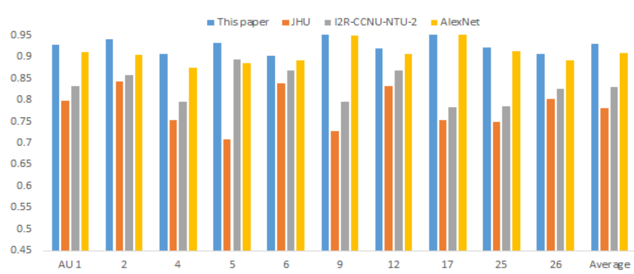

Figure S1. Results on the EmotionNet testing dataset. *Accuracy* calculated using (17).

3. Extended Experimental Results on Landmark detection

Comparative results are given against state-of the-at algorithms and the top performers on the 300-W challenge

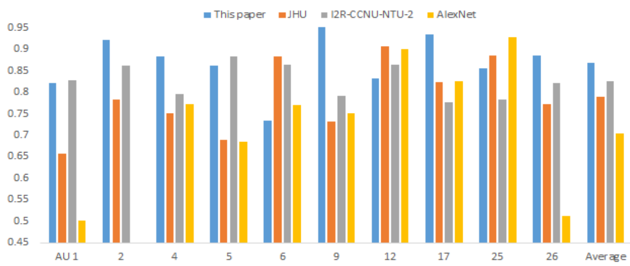

Figure S2. Average $Accuracy$ for images at different scales.

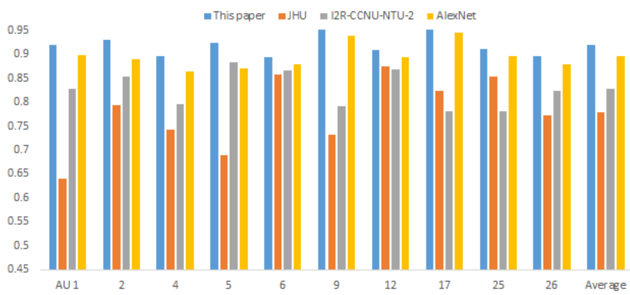

Figure S3. $Accuracy$ for images with small occluders.

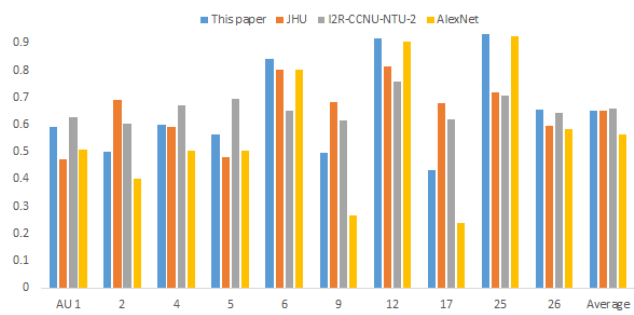

Figure S4. Results on the EmotionNet testing dataset. F_1 score is calculated using (16).

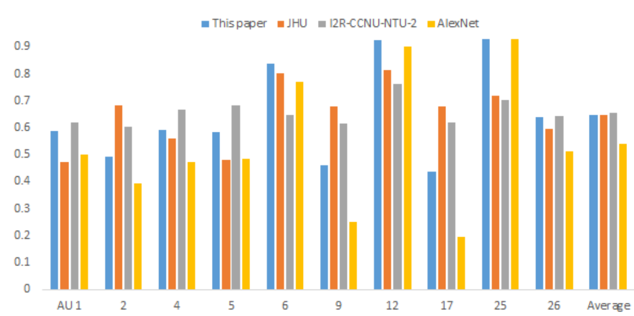

Figure S5. Average F_1 score for images at different scales.

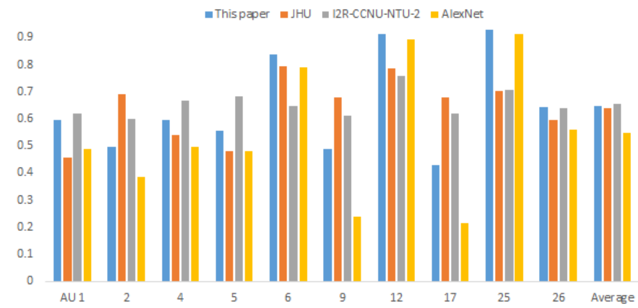

Figure S6. F_1 scores for images with small occluders.

Table S1. Final scores on the EmotionNet Dataset (given by (18)).

AU	This paper	JHU	I2R-CCNU-NTU-2	AlexNet
1	**.76**	.57	.73	.71
2	.72	**.74**	**.73**	.65
4	**.75**	.67	.73	.69
5	.75	.58	**.79**	.7
6	**.87**	.84	.76	.85
9	**.73**	**.72**	.71	.61
12	**.92**	.86	.81	**.91**
17	.7	**.75**	.7	.6
25	**.93**	.8	.75	**.92**
26	**.78**	.69	.74	.74
Average	**.79**	.72	.74	.74

Table S2. F_1 scores on the EmotionNet Dataset (given by (16)).

AU	This paper	JHU	I2R-CCNU-NTU-2	AlexNet
1	.59	.47	**.63**	.51
2	.5	**.69**	.6	.4
4	.6	.59	**.67**	.51
5	.56	.48	**.69**	.5
6	**.84**	.8	.65	.8
9	.5	**.68**	.62	.27
12	**.92**	.81	.76	.9
17	.43	**.68**	.62	.24
25	**.93**	.72	.71	**.93**
26	**.66**	.6	**.65**	.58
Average	**.65**	**.65**	**.66**	.56

dataset [1]. This dataset includes a large number of image variations and a diverse group of people of distinct ethnic and cultural backgrounds. The images are divided following the protocol of [2]: 3,148 images are use for training (to which we apply our data augmentation approach defined above) and 689 faces serve as the testing set. We compare our results with top performing algorithms: Explicit Shape Regression (ESR) [3], Supervised Descent Method (SDM) [4] and Local Binary Features (LBF) [2].

The detection error is the point-to-point squared Euclidean distance, $\|\mathbf{f}_i - \mathbf{y}_i\|_2^2$, normalized by the Euclidean distance between the outer corners of the eyes.

As shown in Table S5, our proposed global method achieves the smallest error. Additionally, Figure S7 compares the results of the proposed GL-CNN algorithm against the top performers in the 300-W challenge [5, 6]. As seen in

Table S3. Average scores (given by (16)) for images at different scales.

AU	This paper	JHU	I2R-CCNU-NTU-2	AlexNet
1	**.71**	.57	**.72**	.65
2	.71	**.73**	**.73**	.59
4	**.74**	.66	**.73**	.68
5	.72	.58	**.78**	.63
6	.79	**.84**	.76	.8
9	**.71**	**.7**	**.7**	.58
12	**.88**	.86	.82	**.88**
17	.69	**.75**	.7	.55
25	**.89**	.8	.74	.86
26	**.76**	.69	.73	.71
Average	**.76**	.72	.74	.69

Table S4. Final scores (equation (18)) for images with small occluders.

AU	This paper	JHU	I2R-CCNU-NTU-2	AlexNet
1	**.76**	.55	.72	.69
2	.71	**.74**	**.73**	.64
4	**.75**	.64	.73	.68
5	.74	.58	**.78**	.68
6	**.87**	.83	.76	.84
9	**.72**	.7	.7	.59
12	**.91**	.83	.81	.89
17	.69	**.75**	.7	.58
25	**.92**	.78	.75	**.91**
26	**.77**	.69	.73	.72
Average	**.78**	.71	.74	.72

Method	Mean normalized error
ESR	7.58
SDM	7.52
LBF	6.32
GL-CNN	**5.47**

Table S5. Average mean normalized error on the 300-W challenge dataset.

this plot, the proposed approach yields results comparable or slightly better than these algorithms but with the added advantage that our method can run at >60 frames/s in Matlab on a i7 desktop.

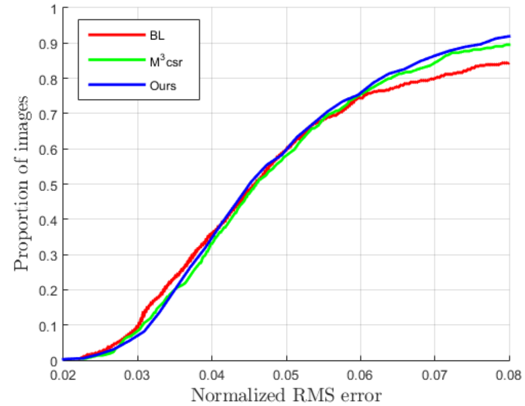

Figure S7. Cumulative normalized root mean square (RMS) error for the top performers on the 300-W challenge [5, 6] and our proposed GL-CNN algorithm. The y-axis specifies the proportion of images, with 1 indicating all images in the database are included. Note that the proposed algorithm outperforms the others when all images are included; demonstrating not only good local fits, but global (overall) fits as well.

References

[1] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "A semi-automatic methodology for facial landmark annotation," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR-W), 5th Workshop on Analysis and Modeling of Faces and Gestures*, 2013. 2

[2] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1685–1692, 2014. 2

[3] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *International Journal of Computer Vision*, vol. 107, no. 2, pp. 177–190, 2014. 2

[4] X. Xiong and F. De La Torre, "Supervised descent method and its applications to face alignment," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 532–539, 2013. 2

[5] J. Deng, Q. Liu, J. Yang, and D. Tao, "M^3 csr: Multi-view, multi-scale and multi-component cascade shape regression," *Image and Vision Computing*, 2015. 2, 3

[6] H. Fan and E. Zhou, "Approaching human level facial landmark localization by deep learning," *Image and Vision Computing*, 2015. 2, 3