Unmasking the Devil in the Details: What Works for Deep Facial Action Coding?

Koichiro Niinuma[1]
kniinuma@us.fujitsu.com
Laszlo A. Jeni[2]
laszlojeni@cmu.edu
Itir Onal Ertugrul[2]
iertugru@andrew.cmu.edu
Jeffrey F. Cohn[3]
jeffcohn@pitt.edu

[1] Fujitsu Laboratories of America
Pittsburgh, PA, USA
[2] Robotics Institute
Carnegie Mellon University
Pittsburgh, PA, USA
[3] Department of Psychology
University of Pittsburgh
Pittsburgh, PA, USA

Abstract

The performance of automated facial expression coding has improving steadily as evidenced by results of the latest Facial Expression Recognition and Analysis (FERA 2017) Challenge. Advances in deep learning techniques have been key to this success. Yet the contribution of critical design choices remains largely unknown. Using the FERA 2017 database, we systematically evaluated design choices in pre-training, feature alignment, model size selection, and optimizer details. Our findings vary from the counter-intuitive (e.g., generic pre-training outperformed face-specific models) to best practices in tuning optimizers. Informed by what we found, we developed an architecture that exceeded state-of-the-art on FERA 2017. We achieved a 3.5% increase in F_1 score for occurrence detection and a 5.8% increase in ICC for intensity estimation.

1 Introduction

In the last-half decade Automated Facial Affect Recognition (AFAR) systems [5] have made major advances in detection of the occurrence and intensity of facial actions, while moving away from controlled laboratory conditions to unconstrained in-the-wild scenarios (see, e.g., [5] and references therein). The evolution of the approaches competing in the previous Facial Expression Recognition & Analysis Challenge (FERA 2015 [26], FERA 2017 [27]) illustrates a shift in the design choices of face representation: data driven deep learning methods are favored over "hand-crafted" shallow representations. In 2015 only a single deep learning method [11] entered the Challenge. It ranked 3rd and 4th in occurrence and intensity detection, respectively. Two years later, deep learning based approaches have dominated the medal podium by a large margin [24, 30].

While the advantage of modern deep learning techniques is clear, little is known about critical design choices among them. Most studies use ad-hoc or default parameters provided by the deep learning frameworks; they neglect to investigate the effect of different parameter settings on facial action unit (AU) detection. Little is known about the relative contribution of

different design choices in pre-training, feature alignment, model size, and optimizer details. A related question is whether a system that exhibits superior performance in a domain in which it has been trained and initially tested will be top performer in another domain[6]. A system may achieve top performance in one domain only to struggle within another domain.

To address questions in design choices, we systematically explored the combinations of different components and their parameters in a modern deep learning based pipeline. Choices included: pre-training practices, image alignment for pre-processing, training set sizes, optimizers, and learning rates (see the different design choices of the current methods in Table 1). By utilizing all the insights, we achieved state-of-the-art performance on both the occurrence and the intensity sub-challenges of FERA 2017 [27] and state-of-the art in cross-domain generalizability to the Denver Intensity of Spontaneous Facial Action (DISFA) dataset [19].

Table 1: An overview of the design choices from studies reporting performance on the FERA 2017 sub-challenges. For occurrence detection, F_1 scores are reported. For intensity detection, Intraclass Correlation coefficients (ICC) are reported. N/A denotes not applicable; N/R denotes not reported. Best scores are denoted in bold.

	Normalization	Architecture	Pre-training	Training set size per model	Optimizer	Learning rate	Occurrence performance (F_1 score)	Intensity performance (ICC)
Valstar et al. [27]	Facial landmarks	Shallow	n/a	n/r	n/a	n/a	0.452	0.217
Li et al. [17]	Facial landmarks	Hybrid	VGG-Face[1]	26,582	n/a	n/a	0.498	n/a
Batista et al. [6]	Face position	Deep	none[2]	1,321,472	Adam	10^{-3}	0.506	0.399
He et al. [12]	Resizing[3]	Hybrid	none	146,847	n/r	n/r	0.507	n/a
Tang et al. [24]	Face position[4]	Deep	VGG-Face	440,541 $+\alpha$[5]	SGD	10^{-3}	**0.574**	n/a
Ertugrul et al. [8]	Face position	Deep	none	1,321,623	Adam	10^{-3}	0.525	n/a
Li et al. [16]	Facial landmarks	Deep	ImageNet-VGG-VD19	260,000 $+\alpha$[6]	SGD	10^{-4}	n/a[7]	n/a
Amirian et al. [1]	Facial landmarks	Shallow	n/a	n/r	n/a	n/a	n/a	0.295
Zhou et al. [30]	Resizing	Deep	ImageNet-VGG-VD16	54,000	SGD	10^{-4}	n/a	**0.446**

[1] A VGG pre-trained model was used to extract features, but not used for classification.
[2] A VGG pre-trained model was used to detect faces, but not used for classification.
[3] Face detection was used for train and validation partition, but not for test partition.
[4] Face position was not directly used, but facial images were cropped by using morphology operations including binary segmentation, connected components labeling and region boundaries extraction.
[5] After down sampling to 440,541 images, Tang et al. increased the number of samples to balance positive and negative samples.
[6] Li et al. increased the number of samples to balance positive and negative samples.
[7] In their paper Li et al. reported F1 scores only on validation partition.

2 Related Work

Numerous approaches have been proposed for action unit (AU) analysis (see [5, 7, 18] and references therein). For most of these, face orientation has been relatively frontal. Where moderate to large non-frontal pose has been considered [13, 15, 21, 23, 25], the lack of a

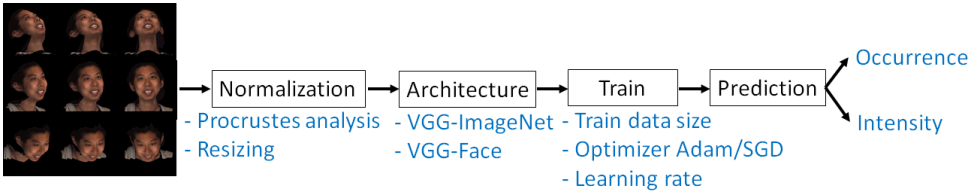

Figure 1: An overview of the experimental design. Blue color denotes design choices and parameters for systematic evaluation.

common protocol has undermined comparisons.

The FERA 2017 Challenge [27] was the first to provide a common protocol with which to compare approaches to detection of AU occurrence and AU intensity robust to pose variation. FERA 2017 provided synthesized face images with 9 head poses as shown in Fig. 1. The training set is based on the BP4D database [28], which includes digital videos of 41 participants. The development and test sets are derived from BP4D+ [29] and include digital videos of 20 and 30 participants, respectively. FERA 2017 presented two sub-challenges: occurrence detection and intensity estimation. For the former 10 AUs were labelled; for the latter, 7 AUs were labelled.

For FERA 2017, the participants proposed a wide range of methods. Table 1 compares them with each other and with two more recent studies from Ertugrul et al.[8] and Li et al.[16]. F_1 score and Intraclass Correlation (ICC) were used to evaluate performance for occurrence detection and intensity estimation, respectively.

Several comparsions are noteworthy. While detailed face alignment using facial landmarks was used for shallow approaches, simple face alignment using face position or resized images more often sufficed for deep learning (DL) approaches. As for architecture, DL performed better than shallow approaches, and DL approaches with pre-trained models performed better than ones without pre-trained model. For both of the sub-challenges, the methods showing the best performance (Tang et al. [24] for occurrence detection, and Zhou et al. [30] for intensity estimation) used DL with a pre-trained model. As for training set size, each method used different number of training images. Adam and SGD were popular choice for optimizer and learning rate varied between 10^{-3} and 10^{-4}.

According to the comparison of the existing methods, the effectiveness of DL approaches, especially the ones using pre-trained models, is indicated for this task, but every approach used a different fixed configuration and the key parameters are unknown. The aim of this study is to investigate the key parameters for both AU occurrence and intensity estimation for this task, and show the optimal configuration.

3 Experiments

The main goal of this study is to investigate the effect of the different components and parameters, and to provide best practices that researchers can use for training deep learning methods for automatic facial expression analysis. Fig. 1 shows an outline of our experimental design. We systematically varied parameters and design choices in this pipeline (key elements are denoted in blue color in Fig. 1).

In every experiments, we explored the effect of optimizer choice and parametric variation of an additional key parameter (image normalization, pre-training choice, training set size, and learning rates). In our baseline configuration we used Procrustes analysis for face alignment and VGG16 network trained on ImageNet. For optimizers, we compared Adaptive Moment Estimation (Adam) and Stochastic Gradient Descent (SGD), with defaults learning rates of 5×10^{-5} and 5×10^{-3}, respectively. We fine-tuned the network from the third convolutional layer using 5,000 images per each pose and AU. Dropout rate was 0.5 throughout of the experiments.

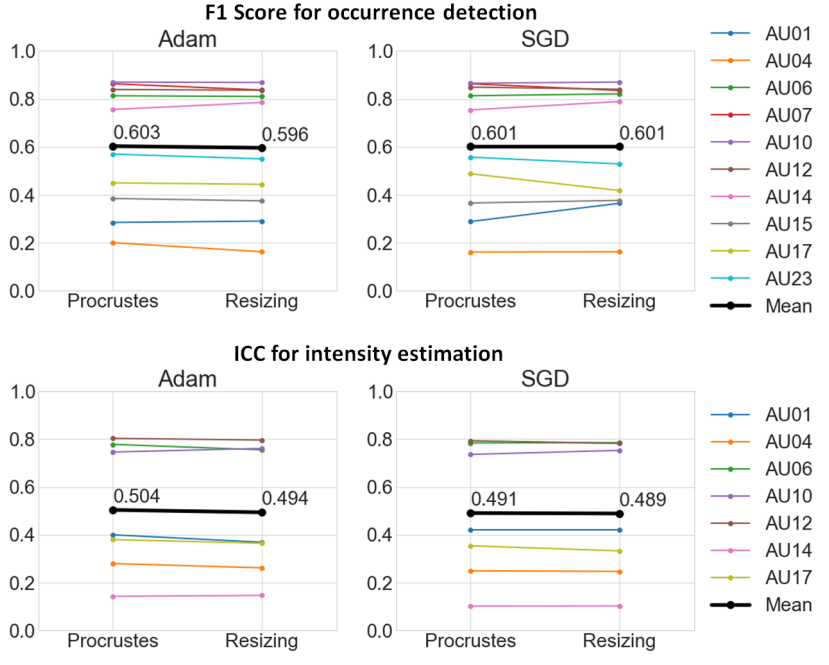

Figure 2: Results on FERA 2017 Test partition with two normalization methods.

3.1 Normalization

We evaluated two methods for image normalization. In the first case we applied Procrustes analysis [10] to the face shapes defined by the landmarks to estimate similarity normalized shapes. In the second case we resized the images to the receptive field of the deep network.

Similarity normalization between source and template shapes using eye locations is a popular choice in the literature. One shortcoming of this approach is that alignment error increases for landmarks farther away from the eye region. This artifact is more prominent under moderate to large head pose variations. To alleviate this problem we used all the 68 landmarks provided by the dlib face tracker[14] to calculate a Procrustes transformation between the predicted shape and a frontal looking template. We chose the size of the template to cover a bounding box of 224x224 pixels, which corresponds to the receptive field of the VGG network.

As for the second option, we resized each input image from the dataset to 224x224 pixel size to match the receptive field of the VGG network.

Fig. 2 shows the F_1 scores and ICC averages for all nine poses for each AU. The left figures show results for Adam optimizer, and the right figures show results for SGD optimizer. The results indicate that the performance with Procrustes analysis is slightly better than the one with Resizing, but the difference is small, only 1%. One possible explanation for this is that the network has enough capacity to learn all the nine different poses present in the training set. Other studies indicate that a form of normalization is often helpful when classifiers are evaluated on poses different from the ones it was trained on [8].

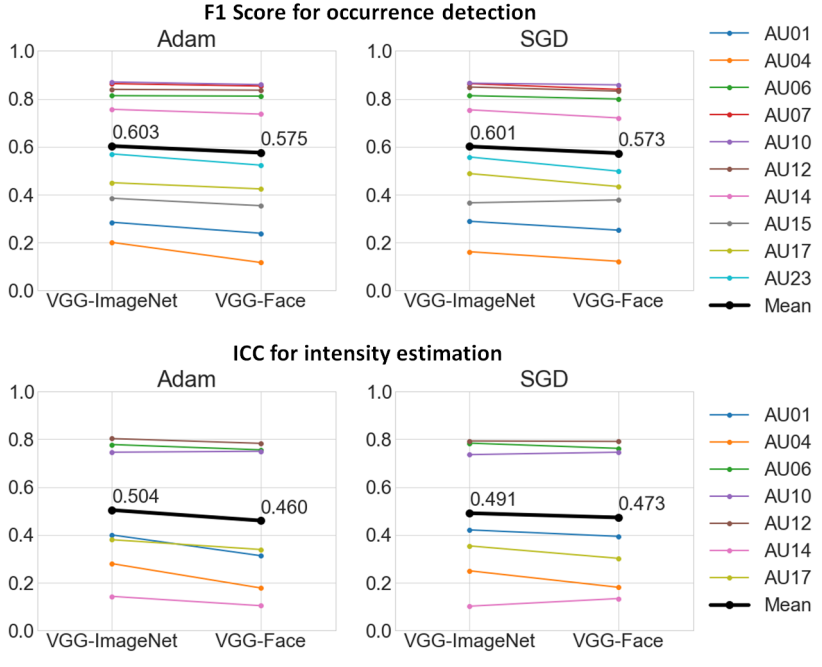

Figure 3: Results on FERA 2017 Test partition with two pre-trained architecture.

3.2 Pre-trained architecture

Training deep models from scratch is time-consuming and the amount of training data at hand may impede good performance. One popular solution is to select a model that was trained on large scale benchmark datasets (source domain) and fine-tune it on the data of our interest (target domain). Although this practice is effective, it is relatively neglected how the type of data in the source domain influence the performance of fine-tuning in the target domain.

To explore this question we selected two models that were trained on very different domains: VGG-16 trained on ImageNet [22] and VGG-Face [20]. We replaced the final layers of each networks with a 2-length one-hot representation for AU occurrence detection, and with a 6-length one-hot representation for the intensity estimation task. In both cases we

trained separate models for each AU, resulting 10 and 7 models for AU occurrence detection and AU intensity estimation, respectively. We fine-tuned our models for 10 epochs and validated performance on the validation partition, then reported results on the subject-independent test partition. We used a PyTorch implementation for all of our models.

Fig. 3 shows that models pre-trained on ImageNet show better performance than the VGG-Face ones. This result seems counter-intuitive since VGG-Face was trained on face images covering a large population, while ImageNet includes many non-face images. One possible explanation is that head-pose variation was small in the data VGG-Face was trained on. In this case a generic image representation is more suitable for the task.

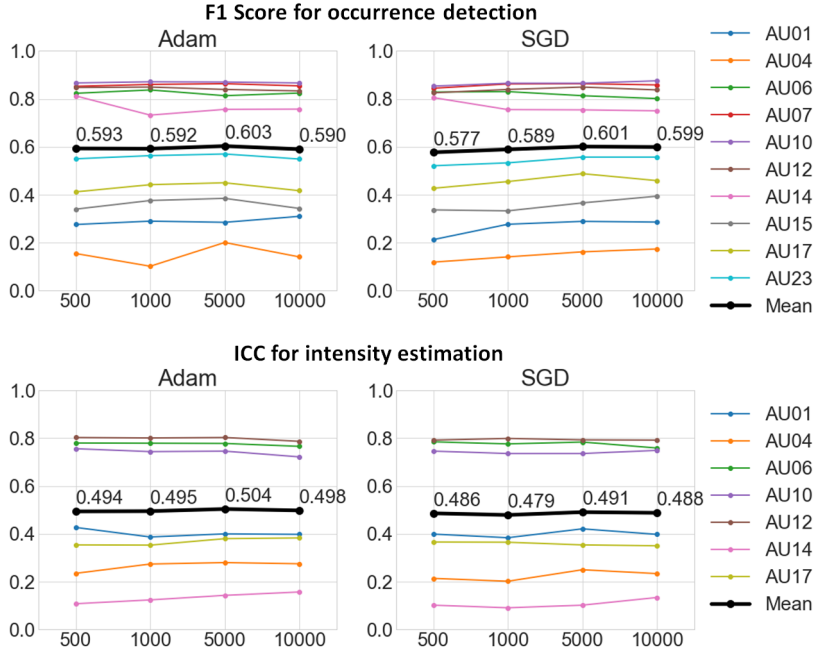

Figure 4: Results on FERA 2017 Test partition with different number of train set size.

3.3 Training set size

Recently, multi-label stratified sampling was found advantageous over naive sampling strategies for AU detection [4]. In this experiment we employ this strategy, and investigate the effect of different training set sizes on the performance. We down-sampled the majority class and up-sampled the minority class to build a stratified training set. We used this procedure for each pose and each AU. For example, in the case of AU occurrence detection, a 5,000 training set size indicate that 5,000 frames with AU present and 5,000 frames where the AU is not present were randomly selected for each pose and for each AU, resulting in 90,000 images in total (=5,000 images x 2 classes x 9 poses).

We repeated the same stratifying procedure with the six ordinal classes of the intensity sub-challenge. In this case, a 5,000 training set size means that 5,000 images were randomly

selected from the six classes (not present, and A to E levels) for each pose and for each AU, resulting in 270,000 images in total (=5,000 images x 6 classes x 9 poses).

Fig. 4 shows results with as the function of different training set size. The training set size have minor influence on the performance: scores peaked at 5,000 images, after that performance plateaued.

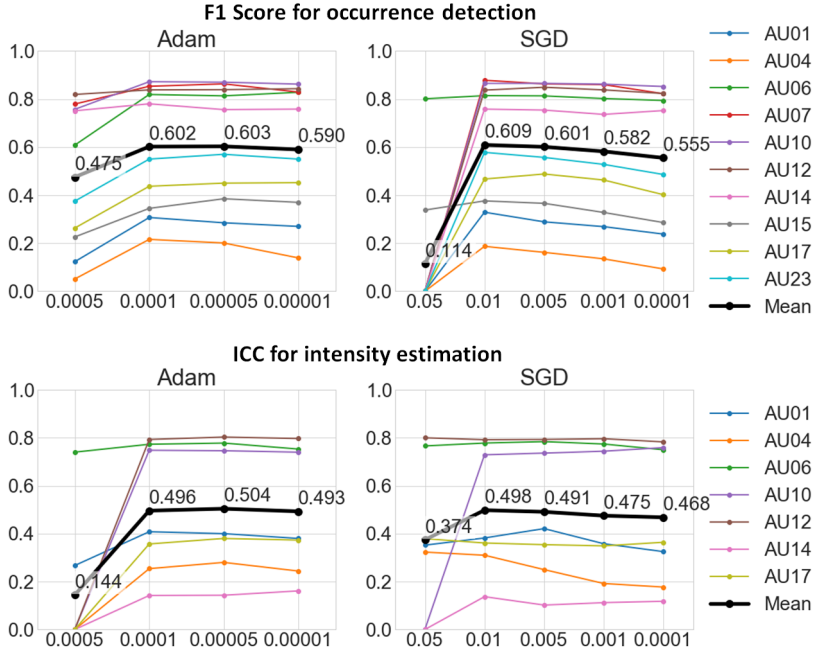

Figure 5: Effect of learning rates and choice of optimizers on the FERA 2017 Test partition.

3.4 Optimizer and learning rate

In this experiment we investigated the impact of different optimizers and learning rates (LR) on the performance. We varied the learning rates, but other optimizer parameters were set to the default values used in PyTorch: betas=(0.9, 0.999) without weight decay for Adam, and no momentum, no dampening, no weight decay and no Nesterov acceleration for SGD.

Fig. 5 shows that the optimal learning rate depends on the choice of optimizer. For Adam, LR=5×10^{-5} gave the best results, and for SGD, LR=0.01 reached the best performance for both occurrence detection and intensity estimation. In addition, we can see that the performance differences between Adam and SGD are negligible if one uses the optimal learning rates for each optimizer, respectively.

It is worth noting that Zhou et al. [30] used SGD with LR=10^{-4} for the AU intensity estimation task. Our results indicate that using Adam optimizer or SGD optimizer with larger learning rate could have improved their performance. Tang et al. [24] used SGD with LR=10^{-3}, but they also applied momentum. Our additional experiments revealed that when momentum is used for SGD, smaller learning rate is preferable for optimal performance.

More specifically, when we used the same parameters as Tang et al. [24] reported for SGD (momentum=0.9, weight decay=0.02) F_1 score peaked at 0.596 using LR=10^{-4}. Their learning rate is close to optimal, though SGD without momentum further improves F_1 score to 0.609 with LR=0.01.

We note that when the learning rate was set to a large value some models did not converge and predicted the majority class for all samples. Under this rare condition ICC converges to zeros, but this should not be interpreted as chance performance. As variation in predicted intensity values reduces, the ICC metric loses predictive power.

3.5 Comparison with existing methods

We compare our method with the state-of-the-art on both the AU occurrence detection (Table 2) and the AU intensity estimation (Table 3) sub-challenges from FERA 2017. The final parameters of our models are nearly identical for the two tasks: we used face alignment with Procrustes analysis as a pre-processing step, and we fine-tuned ImageNet pre-trained VGG16 model on stratified sets consisting of 5,000 samples per each class, pose, and AU.

For AU occurrence detection, SGD with LR=0.01 gave the best result ($F_1 = 0.609$), while for AU intensity estimation, Adam with LR=5×10^{-5} reached the best performance (ICC = 0.504). These scores outperform other state-of-the-art methods.

We note a few key differences that contributed to this achievement. The main difference with Tang et al. [24] is that they used VGG-Face pre-trained model while we used ImageNet pre-trained model. Zhou et al. [30] used used SGD with small learning rate while the combination of our optimizer and learning rate is optimal. While Li et al. [16] evaluated their method for AU occurrence detection using the FERA 2017 dataset, they report performance only on the Validation partition. Their best F_1 score (0.522) is 9% lower than ours (0.611) on the Validation partition.

Table 2: F1 scores for occurrence detection results on FERA 2017 Test partition.

	Valstar et al. [27]	Li et al. [17]	Batista et al. [3]	He et al. [12]	Ertugrul et al. [8]	Tang et al. [24]	Our model
AU1	0.147	0.215	0.219	0.198	0.196	0.263	**0.329**
AU4	0.044	0.044	0.056	0.043	0.067	0.118	**0.187**
AU6	0.630	0.755	0.785	0.747	0.766	0.776	**0.814**
AU7	0.755	0.805	0.816	0.784	0.791	0.808	**0.878**
AU10	0.758	0.810	0.838	0.816	0.840	**0.865**	**0.865**
AU12	0.687	0.753	0.780	0.809	0.819	**0.843**	0.837
AU14	0.668	0.750	0.747	0.691	**0.764**	0.757	0.758
AU15	0.220	0.208	0.145	0.208	0.247	0.362	**0.376**
AU17	0.274	0.286	0.388	0.398	0.349	0.424	**0.467**
AU23	0.342	0.356	0.286	0.374	0.413	0.519	**0.578**
Mean	0.452	0.498	0.506	0.507	0.525	0.574	**0.609**

3.6 Cross-domain evaluation

Differences in illumination, cameras, orientation of the face, quality and diversity of the training data influence predictive performance between domains. To evaluate the generalizability of our method to unseen conditions, we report performance on the Denver Intensity of Spontaneous Facial Action (DISFA) [19] dataset.

Table 3: ICC for intensity estimation on FERA 2017 Test partition.

	Valstar et al. [27]	Amirian et al. [1]	Batista et al. [3]	Zhou et al. [30]	Our model
AU1	0.035	0.169	0.228	0.307	**0.400**
AU4	-0.004	0.021	0.057	0.147	**0.280**
AU6	0.461	0.509	0.702	0.671	**0.778**
AU10	0.451	0.590	0.710	0.735	**0.746**
AU12	0.518	0.615	0.732	0.793	**0.803**
AU14	0.037	-0.027	0.104	**0.147**	0.143
AU17	0.020	0.190	0.260	0.319	**0.380**
Mean	0.217	0.295	0.399	0.446	**0.504**

DISFA dataset was annotated with AU intensity labels. To create binary AU occurrences, we thresholded the 6-points intensity values at A-level (A-level or higher means the AU is present). We evaluate both occurrence detection and intensity estimation performance of our system.

In these experiments, we used the previously trained CNN models reported in Section 3.5. No fine tuning was performed on the target domain. We used the built-in face detector in dlib [14] to detect the face before applying Procrustes analysis.

Both Ghosh et al. [9] and Baltrušaitis et al. [2] used BP4D to train their model, and thresholded AU intensity values at A-level to create binary events. For a fair comparison, we also report Accuracy and 2AFC scores, that Ghosh et al. [9] used. We outperform their method in both metrics. Baltrušaitis et al. [2] report cross-domain scores only for two AUs (AU 12 and AU17). Our models show better performance for both cases. These results show the robustness of our model for cross-domain situation.

Table 4: Comparison of cross-domain performance to DISFA dataset for occurrence detection and intensity estimation.

	Occurrence Detection						Intensity Estimation
	Accuracy		2AFC		F1		ICC
	Our	[9]	Our	[9]	Our	[2]	Our
AU01	**0.932**	0.838	**0.714**	0.660	0.475	-	0.533
AU04	0.806	**0.833**	0.723	**0.740**	0.531	-	0.560
AU06	**0.860**	0.703	0.758	**0.870**	0.567	-	0.451
AU12	**0.859**	0.624	0.859	**0.873**	**0.742**	0.700	0.747
AU15	**0.823**	0.752	**0.671**	0.617	0.253	-	-
AU17	**0.738**	0.689	**0.742**	0.585	**0.361**	0.260	0.319
Mean	**0.836**	0.740	**0.745**	0.724	**0.488**	-	0.522

4 Conclusions

We addressed how design choices influence performance in facial AU coding using deep learning systems, by evaluating the combinations of different components and their parameters present in such systems.

We found that the source domain in which pre-training was performed influenced the performance of fine-tuning in the target domain. Counter-intuitively, generic pre-training proved better, than a face specific one. Another important factor contributing to the performance is the choice of different learning rates for different optimizers. We found that Adam optimizer with small learning rate and SGD with large learning rate is optimal for expression coding. Best parameters of the optimizers were similar for both AU occurrence and intensity estimation, while varying the training set size and the type of image normalization had little effect on performance.

5 Acknowledgments

This research was supported in part by Fujitsu Laboratories of America, NIH awards NS100549 and MH096951, and NSF award CNS-1629716.

References

[1] M. Amirian, M. Kächele, G. Palm, and F. Schwenker. Support vector regression of sparse dictionary-based features for view-independent action unit intensity estimation. *In Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 854–859, 2017.

[2] T. Baltrušaitis, M. Mahmoud, and P. Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. *In Automatic Face & Gesture Recognition and Workshops (FG 2015), 2015 IEEE International Conference on*, 2015.

[3] J. C. Batista, V. Albiero, O. R. P. Bellon, and L. Silva. AUMPNet: simultaneous action units detection and intensity estimation on multipose facial images using a single convolutional neural network. *In Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 868–871, 2017.

[4] W.-S. Chu, F. De la Torre, and J. F. Cohn. Learning facial action units with spatiotemporal cues and multi-label sampling. *Image and vision computing*, 81:1–14, 2019.

[5] J. F. Cohn and F. De la Torre. Automated face analysis for affective. In *The Oxford handbook of affective computing*, page 131. 2014.

[6] J. F. Cohn, I. O. Ertugrul, W.-S. Chu, J. M. Girard, L. A. Jeni, and Z. Hammal. Chapter 19 - affective facial computing: Generalizability across domains. In *Multimodal Behavior Analysis in the Wild*, Computer Vision and Pattern Recognition, pages 407 – 441. Academic Press, 2019.

[7] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero. Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1548–1568, 2016.

[8] I. O. Ertugrul, L. A. Jeni, and J. F. Cohn. FACSCaps: Pose-independent facial action coding with capsules. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2130–2139, 2018.

[9] S. Ghosh, E. Laksana, S. Scherer, and L.-P. Morency. A multi-label convolutional neural network approach to cross-domain action unit detection. *International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2015.

[10] J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

[11] A. Gudi, H. E. Tasli, T. M. Den 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)*, volume 6, pages 1–5, 2015.

[12] J. He, D. Li, B. Yang, S. Cao, B. Sun, and L. Yu. Multi view facial action unit detection based on CNN and BLSTM-RNN. *In Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 848–853, 2017.

[13] L. A. Jeni, A. Lorincz, T. Nagy, Zs. Palotai, J. Sebok, Z. Szabo, and D. Takacs. 3D shape estimation in video sequences provides high precision evaluation of facial expressions. *Image and Vision Computing*, 30(10):785 – 795, 2012.

[14] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[15] S. Kumano, K. Otsuka, J. Yamato, E. Maeda, and Y. Sato. Pose-invariant facial expression recognition using variable-intensity templates. *International Journal of Computer Vision*, 83(2):178–194, 2009.

[16] W. Li, F. Abtahi, Z. Zhu, and L. Yin. EAC-Net: Deep nets with enhancing and cropping for facial action unit detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(11):2583–2596, 2018.

[17] X. Li, S. Chen, and Q. Jin. Facial action units detection with multi-features and -AUs fusion. *In Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 860–865, 2017.

[18] B. Martinez, M. F. Valster, B. Jiang, and M. Pantic. Automatic analysis of facial actions: A survey. *IEEE Transactions on Affective Computing*, 2017.

[19] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. DISFA : A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4 (2):151–160, 2013.

[20] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. *British Machine Vision Conference*, 2015.

[21] O. Rudovic, M. Pantic, and I. Patras. Coupled gaussian processes for pose-invariant facial expression recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(6):1357–1369, 2013.

[22] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *In International Conference on Learning Representations (ICLR)*, 2015.

[23] S. Taheri, P. Turaga, and R. Chellappa. Towards view-invariant expression analysis using analytic shape manifolds. In *Face and Gesture 2011*, pages 306–313, 2011.

[24] C. Tang, W. Zheng, J. Yan, Q. Li, Y. Li, T. Zhang, and Z. Cui. View-independent facial action unit detection. *In Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 878–882, 2017.

[25] 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*, pages 359–371, 2016.

[26] M. F. Valstar, T. Almaev, J. M. Girard, G. McKeown, M. Mehu, L. Yin, M. Pantic, and J. F. Cohn. FERA 2015-second facial expression recognition and analysis challenge. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 6, pages 1–8, 2015.

[27] M. F. Valstar, E. Sánchez-Lozano, J. F. Cohn, L. A. Jeni, J. M. Girard, Z. Zhang, L. Yin, and M. Pantic. FERA 2017 - addressing head pose in the third facial expression recognition and analysis challenge. *In Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 839–847, 2017.

[28] X. Zhang, L. Yin, J. F. Cohn, S Canavan, M. Reale, A. Horowitz, and J. M. Girard. BP4D-spontaneous: a high-resolution spontaneous 3D dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.

[29] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, J. Cohn, Q. Ji, and L. Yin. Multimodal spontaneous emotion corpus for human behavior analysis. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3438–3446, 2016.

[30] Y. Zhou, J. Pi, and B. E. Shi. Pose-independent facial action unit intensity regression based on multi-task deep transfer learning. *In Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 872–877, 2017.