JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 09, ISSUE 04, 2022

A Combinatorial Review of Emotion Recognition As Rendered By Specialized Convolutional Neural Networks

F. LUDYMA FERNANDO [1], S. JOHN PETER[2]

[1][2]Manonmaniam Sundaranar University, Tirunelveli

Corresponding Author:

F. Ludyma Fernando

34, Kerecope Street, Thoothukudi – 628001,ludy0928@gmail.com

Abstract:

Recognition of a set of patterns and classifying them to a predefined set of standards are not new to the field of Artificial Intelligence, of which the process called Data Mining plays an extremely important part. Having an idea of what actually needs to be done while recognizing emotions from a dataset, I have reviewed a few methods that have been used to detect emotions and their accuracy and have tried to conclude the ideal method for recognizing emotions and the improvements that can be made for the same to increase its accuracy. The papers taken into account are the ***Dominant and Complementary Emotion Recognition from the still images of faces***[1] and ***Emotion Recognition in Context*** [2]. Both the authors have tried their best to prove the efficiency of their research. The first paper analyses various methods and concludes the best of a set of three methods. The second paper gives one concrete solution to effectively identify a set of emotion categories.

Keywords: Emotion Recognition, Convolutional Neural Networks(CNN), Emotions

1. Introduction:

The process of recognition of emotions from a dataset of images also comes under the category. Emotions are broadly defined into 6 major classes – happiness, fear, anger, sadness, surprise and disgust[1]. They may be further classified into negative as well as positive emotions or they can be further combined into compound emotions like happily surprised. The emotions may be discrete or continuous. Whatever the kind of emotion, it needs to be recognized from a set of images from a dataset that is exclusively used for the purpose. The larger the amount of data available, the higher the probability of accuracy in retrieving emotions. The images are trained and validated with the structures normally used in data mining like the Neural Networks, Convolutional Neural Networks, Support Vector Machines etc. The factors that are to be considered while differentiating the emotions from the other on the dataset may be geometrical or textural. Whatever the structure used to recognize emotions, the part of the input image that has to be processed is crucial.

2. Literature Review:

The first paper considered, is the *Dominant and Complementary Emotion Recognition from still images of faces*[2]. This paper has given a detailed view on the dominant emotions such

ISSN- 2394-5125 VOL 09, ISSUE 04, 2022

as happiness, sadness etc and also the complementary emotions such as happily surprised, fearlessly happy etc. The methods that have been used to detect those emotions and its accuracy have also been elaborated. The paper uses the *iCV-MEFED* Dataset[3] to determine dominant and complementary emotions. It makes use of the past works published earlier to find the most accurate way of calculating emotions. The authors have investigated four of the previous works which included those of Du et al[4], Alex & Du[5], Benitwz Quiroz[6] and Le and Ding[7]. Aggregately they have found out that the methods of detecting emotions through *facial landmarks (geographical)* are best for recognizing emotions rather than using spatial textural features. They had used different datasets which was used as training sets to detect the basic emotions as well as the compound facial expressions. The average maximum accuracy as found out by the four authors is 81% which is still much a decent rate for precise emotion detection. They have also taken up three methods which mainly uses Convolutional Neural Network (CNN) and have concluded with the 'winner' method, which gives accuracy higher than the methods discussed earlier. The methods discussed are as follows:

I. Multi-modality network with visual and Geometrical Information:

This method used Dlib[8] library for facial landmark and face alignment. Then facial landmarks are refined after facial alignment. Each face *i* is represented by an average *lm* landmark face:

$$lm^{(i)} = \frac{1}{N} \sum_{j=1}^{N} l_N^{(i)}$$

where N is the number of samples which is about 250 in iCV-MEDFED dataset and l is the flattened vector of landmark. Finally, the geometrical representation extracted as the landmark displacement:

$$lr^{(i)} = l^{(i)} - lm^{(i)}$$

lr is the landmark residual (or displacement)

The Network structure consists of both textural and geometrical features represented by $p1 \in R^{256}$ and $p2 \in R^{136}$ respectively which are then concatenated to $p \in R^{392}$ and fed into a fully connected layer. Some samples can be correctly classified by this p1 and p2 which spans vectors and is done by mapping vectors of lower dimensions to higher dimensions. The method uses stochastic gradient descent with a mni batch size of 32 and a max iteration is 1×10^5. The learning rate is divided by 5 in every 20000 iterations. A weight decay of 5 x 10 - 4 and a momentum of 0.9 is adopted. At test stage, p1 and p2 are computed, concatenated and given as input to the classifier.

II. Unsupervised Learning of Convolutional Neural Networks:

Similar to the previous method, this network also extracts and aligns face using DLib library. The images of the faces are resized to 96x96x3. The unsupervised learning model [8] used here is the CNN model with filters trained layer-wise using k-means clustering. It is a simple model but a wider shallow network seemed to give improved accuracy than the deeper ones.

ISSN- 2394-5125 VOL 09, ISSUE 04, 2022

The CNN structure consisted of a batch – norm layer, convolutional layer with 1024 filters (15 x 15 x 3), max pooling (12 x 12), a Rectified Linear Unit (ReLU) rectifier and a rootsift normalization. Principal Component Analysis (PCA) is applied to extracted features. A linear SVM is chosen as a classifier and all 50 emotions are treated as independent. Filters are trained using k-means with ZCA whitening. Filter size, polling type and size, as well as SVM regularization constant are selected by 5 fold cross validation. At feature extraction stage, a mini batch size of 25 and augmentation of horizontal flipping are adopted. As different SVMs are employed, based on distinct sets of extracted features, final prediction is obtained by averaging individual SVM scores.

III. Inception- V3 structuring with Auxiliary center loss:

This prediction method uses a Inception – V3 network structure. This paper has adopted center loss[9] as an auxiliary optimization function. The multi task CNN is adopted to parse face bounding boxes and landmarks. Then face images are aligned by affine transformation and resized to 224 x 224 x3. Feature Extraction is done by using Inception – V3 CNN. Finally cross entropy is applied for optimization. This loss enhances the ability of the model to distinguish similar samples and improves overall performance. This network is optimized by the SGD and 87 and the learning rates are fixed as 10^{-3} and 10^{-4}. Weight decay is 4×10^{-4}, momentum is 0.9 and all layers are initialized.

While the three methods have been discussed in detail as above, the experimental analysis has also been made for the same. The analysis included measuring the accuracy on the grounds of overall recognition as well as in each emotion category, confusion matrix analysis and computational cost analysis.

Methods	**Misclassification (Validation set)**	**Misclassification(Test set)**
Multi Modality network with Visual and Geometrical Information	0.793	0.802
Unsupervised Learning of Convolutional Neural Network	0.840	0.853
Inception – V3 structure with auxiliary center loss	0.875	0.877

Table 2.1: The misclassification rates of the three methods

The table shows the accuracy of the three methods as calculated by the misclassification rates of their data sets. We can infer that the Multi modality network with visual and geometrical information stands out from the rest, because unlike its competitors, it uses both the textural as well as the geometrical features.

As far as the classification if emotion categories are concerned, all the three methods have complemented each other. The emotion categories and their values are as follows:

ISSN- 2394-5125 VOL 09, ISSUE 04, 2022

Methods	Values	Emotion Category
Multi Modality network with Visual and Geometrical Information	0	Neutral
	1	Angry
	9	Contempt
	33	Happy
	35	Happily Surprised
	46	Surprisingly fearful
	49	Surprised
Unsupervised Learning of Convolutional Neural Network	7	Angrily surprised
	15	Disgustingly Angry
	29	Happily Angry
	41	Sad
Inception – V3 structure with auxiliary center loss	2	Angrily Contempt
	5	Angrily Happy
	47	Surprisingly Happy

Table 2.2: Accuracy of Each Emotion Category

The table shows the different emotion categories that can be determined using the corresponding methods. It is evident that different emotion categories can be recognized by different methods which can lead to combining the methods for a broader range of emotions..

The confusion matrix for all the three methods prove that it is harder to classify the dominant and the complementary emotions accurately, because the methods sift out the dominant emotions and categorize according to it. For example, when the exact emotion was surprisingly happy, the method recognized it as happily surprised.

The other aspect is about the computation time, which I have ignored for the case of this review.

The second paper that I have taken into consideration is *Emotion Recognition in Context*[10]. This paper considers 26 emotion categories that could be detected along with the continuous dimensions of *valence, arousal and dominance* [11]. The dataset that is being used is the *EMOTIC* (Emotions in Context) database[12][13] which contains images of people in context *under uncontrolled environments.* A Convolutional Neural Network has been trained which jointly analyses the person along with the environment to provide rich information about emotional states.

This paper analyses the 26 emotional states by considering not only the facial features, but also by understanding the body language and the body pose features, that is, its context[14]. This is a new turn in emotion recognition. This is done by taking a wider view of the person. The additional emotions that could be recognized include yearning, self esteem, peace,

ISSN- 2394-5125 VOL 09, ISSUE 04, 2022

annoyance etc. The conclusions made in this research are that the context contributes the relevant information and that combining the emotional categories and the continuous dimensions during training results in a more robust system for recognizing emotional states. The following are the 26 discrete categories of emotions that are taken into account.

1. **Peace**: well being and relaxed; no worry; having positive thoughts or sensations; satisfied
2. **Affection**: fond feelings; love; tenderness
3**. Esteem**: feelings of favorable opinion or judgment; respect; admiration; gratefulness
4. **Anticipation**: state of looking forward; hoping on or getting prepared for possible future events
5**. Engagement**: paying attention to something; absorbed into something; curious; interested
6**. Confidence**: feeling of being certain; conviction that an outcome will be favorable; encouraged; proud
7. **Happiness:** feeling delighted; feeling enjoyment or amusement
8. **Pleasure**: feeling of delight in the senses
9. **Excitement**: feeling enthusiasm; stimulated; energetic
10. **Surprise**: sudden discovery of something unexpected
11.**Sympathy**: state of sharing others emotions, goals or troubles; supportive; compassionate
12**. Doubt/Confusion**: difficulty to understand or decide; thinking about different options
13. **Disconnection**: feeling not interested in the main event of the surrounding; indifferent; bored; distracted
14. **Fatigue**: weariness; tiredness; sleepy
15. **Embarrassment**: feeling ashamed or guilty
16. **Yearning**: strong desire to have something; jealous; envious; lust
17. **Disapproval**: feeling that something is wrong or reprehensible; contempt; hostile
18. **Aversion**: feeling disgust, dislike, repulsion; feeling hate
19. **Annoyance**: bothered by something or someone; irritated; impatient; frustrated
20. **Anger**: intense displeasure or rage; furious; resentful
21. **Sensitivity**: feeling of being physically or emotionally wounded; feeling delicate or vulnerable
22. **Sadness**: feeling unhappy, sorrow, disappointed, or discouraged
23. **Disquietment**: nervous; worried; upset; anxious; tense; pressured; alarmed
24. **Fear**: feeling suspicious or afraid of danger, threat, evil or pain; horror
25. **Pain:** physical suffering
26. **Suffering:** psychological or emotional pain; distressed; anguished

Table 2.3: Emotion Categories

While mentioning about the Continuous dimensions the words that are used to denote the categories of emotions must be determined, which is quite complicated. So, to keep the vocabulary in tune with the project and its outcome, two conditions have been set as the

ISSN- 2394-5125 VOL 09, ISSUE 04, 2022

benchmark for naming emotions.- *Disjointness and Visual Separability*. For example, emotions like, disgust, hate, dislike and repulsion seem to have the same kind of emotion and its difficult to separate visually. Hence there is a need for the conditions. For annotating images according to the specified categories and the continuous dimensions, the *Amazon Mechanical Turk (AMT)* interface was designed. The annotations include the gender and the age of the subjects. The database statistics includes how the emotions are being categorized based on the VAD Emotional State Model[15]. For example,

- The emotions suffering and pain have lowest valence score and the emotions affection and happiness have the highest valence scores.
- The emotions fatigue and sadness have lowest arousal score and the emotions confidence and excitement have the highest arousal scores
- The emotions suffering and pain have lower dominance level(since people feel that they don't have control over the situation) and the emotions confidence and excitement have higher dominance level.

The proposed CNN model has *two feature extractors and a fusion module*. The first module takes the relevant features of the area where the person whose emotion has to be extracted. The second module processes the entire image to get contextual information with regard to the emotions to be detected. Finally the fusion module, which is actually a fusion network, which takes as input the image and the body features and estimates the discrete categories and the continuous dimensions. The feature extraction module is the truncated version of the low-rank filter Convolutional network as proposed in [16]. The original network consisted of 16 convolutional layers with 1-dimensional kernels, effectively modeling 8 2-dimensional kernals. For this proposed network, the truncated version removes the fully connected layer and output the features from the activation map of the last convolutional layer. This was done to localize the different parts of relevance in the image. These features are combined using a fusion network. It uses a *global average pooling layer* on each feature map to reduce the number features. The output of this layer is a 256 dimension vector. Then a large fully connected network is included for the learning process. Separate branches are given for the continuous dimensions and discrete emotion categories. Batch normalization and rectifier linear units have been added at each convolutional layer. The *overall loss* for training the model has been defined as the weighted combination of two individual losses. For both the Continuous dimensions and as well as the discrete emotions, the task is formulated as a regression problem with Euclidean loss to compensate for the class imbalance existing in the dataset. The overall loss for training the model is defined as the weighted combination of two individual losses, as given by,

$$L_{comb} = \lambda_{disc} L_{disc} + \lambda_{cont} L_{cont}$$

where the $\lambda_{disc,cont}$ weights the importance of each loss and $L_{disc}$and L_{cont} are the loss due to learning discrete and continuous categories respectively.

They have been given separately as

ISSN- 2394-5125 VOL 09, ISSUE 04, 2022

$$L_{disc} = \frac{1}{N}\sum_{i=1}^{N} w_i \left(\hat{y}_i^{disc} - y_i^{disc}\right)^2$$

where N is the number of categories (N=26), $\hat{y}_i^{disc}$ is the estimated output for the i-th category and y_i^{disc}, the ground truth label. w_i is the weight assigned to each category.

$$L_{cont} = \frac{1}{\#C}\sum_{k\epsilon C} v_k \,(\hat{y}_k^{cont} - y_k^{cont})^2$$

where C = {Valence, Arousal, Dominance},$\hat{y}_k^{cont}$ and y_k^{cont} are the estimated output and the normalized groundtruth for the k-th dimension and v_k= [0, 1] is a weight to represent the error margin. v_k= 0 if $|\hat{y}_k^{cont} - y_k^{cont}| < \theta$ otherwise v_k= 1.

The following images of their findings have shown the average precision rate and and mean error rate obtained per each dimension for different CNN configurations: Body (B), Image (I) and Body + Image (B+I)

Figure 2.1 : Average Precision Rate

Category	CNN Inputs and Loss			
	B	I	B + I	B + I
	L_{comb}			L_{disc}
1. Peace	20.63	20.43	**22.94**	20.03
2. Affection	21.98	17.74	**26.01**	20.04
3. Esteem	18.83	**19.31**	18.58	18.95
4. Anticipation	54.31	49.06	**58.99**	52.59
5. Engagement	82.17	78.48	**86.27**	80.48
6. Confidence	74.33	65.42	**81.09**	69.17
7. Happiness	54.78	49.32	**55.21**	52.81
8. Pleasure	48.65	45.38	48.65	**49.23**
9. Excitement	74.16	68.82	**78.54**	70.83
10. Surprise	21.95	19.71	**21.96**	20.92
11. Sympathy	11.68	11.30	**15.25**	11.11
12. Doubt/Confusion	33.49	33.25	**33.57**	33.16
13. Disconnection	18.03	16.93	**21.25**	16.25
14. Fatigue	9.53	7.30	**10.31**	7.67
15. Embarrassment	2.26	1.87	**3.08**	1.84
16. Yearning	8.69	7.88	**9.01**	8.42
17. Disapproval	12.32	6.60	**16.28**	10.04
18. Aversion	8.13	3.59	**9.56**	7.81
19. Annoyance	11.62	6.04	**16.39**	11.77
20. Anger	7.93	5.15	**11.29**	8.33
21. Sensitivity	5.86	4.94	**8.94**	4.91
22. Sadness	9.44	6.28	**19.29**	7.26
23. Disquietment	18.75	16.85	**20.13**	18.21
24. Fear	15.73	14.60	**16.44**	15.35
25. Pain	6.02	2.98	**10.00**	4.17
26. Suffering	10.06	5.35	**17.60**	7.42
Average	25.44	22.48	**28.33**	24.18

ISSN- 2394-5125 VOL 09, ISSUE 04, 2022

Figure 2.2 : Mean Error Rate

Dimension	CNN Inputs and Loss			
	B	I	B + I	B + I
	L_{comb}			L_{cont}
Valence	**0.9**	**0.9**	**0.9**	1.0
Arousal	**1.1**	1.9	1.2	1.5
Dominance	1.0	**0.8**	0.9	**0.8**
Average	**1.0**	1.2	**1.0**	1.1

From inferred data it could be clearly seen that in Figure 2.1 and 2.2, the best data was obtained when the Body and Image were used as inputs. Using the validation set the threshold for each category was taken when the value where *Precision = Recall*, to detect each category. Then Jaccard Coefficient have been calculated for each category which is above 0.4 meaning that the significant number of categories were correctly retrieved. Hence the method being effective in combining the body, face and the environmental aspects into account, is able to detect the emotions even when the face is not shown, also when the environments were non – controlled.

3. Author's Review:

The papers clearly show that recognition of emotions is scalable and the methods can be fine tuned to get perfect outcome. The Convolutional Neural Networks has proven to be an efficient structure for recognizing patterns of emotions, even when the emotions are discrete or continuous, though it has been ranked the second in the former paper. The research also concludes that the Activation Functions in the layers can be scrutinized to get more accurate results. From the inferences made in table 2.1, the first method, though it was named to be the most efficient among the three, have recognized less percentage of compound emotions as compared to other two, of which the method which used the Convolutional Neural Network is the second. The same kind of interpretation can be done with the second paper which uses CNN that classifies emotions spanning over 26 emotion categories. When there is a large range of imput to be classified into a smaller range then, the vanishing gradient problem occurs and that is the reason why the ReLU (Rectified Linear Unit) has been used in the first paper. The ReLU activation function can be implemented in CNN which has been structured in [2] to maximize the Average accuracy rate, because ReLU makes sure that all the input are perfectly classified into the smallest range of emotions. It cannot be disagreed that the CNNs that are used, act more effectively when they are shallow than deep. Hence in [2] the proposed CNN has been truncated to give more accurate results. The reason I mention this, is that, the ReLU activation function has the disadvantage of Exploding Gradients and Vanishing Gradients. These setbacks make the values remain constant for higher number of epochs. A better activation function like the SELU (Scaled Exponential Linear Unit) can be used in this case. SELU is a self normalizing function which rules out the Exploding and Vanishing Gradients issue.

Scaled Exponential Linear Unit (SELU)[16]: This activation function is one of the newer functions. The best part about the SELU activation is that of internal normalization by the use of LeCun_Normal and AlphaDropout for weight initialization and application of dropouts

ISSN- 2394-5125 VOL 09, ISSUE 04, 2022

respectively. The authors of the SELU activation have calculated two values: alpha (α) nad lambda (λ) value for the equation. The values are:

$\alpha \approx 1.6732632423543772848170429916 71$

$\lambda \approx 1.0507009873554804934193349852 94$

The equation for the SELU activation:

$$\text{SELU}(x) = \lambda \begin{cases} x & \text{if } x > 0 \\ ae^x - \alpha & \text{if } x \leq 0 \end{cases}$$

(i.e) if the input value is greater than zero, the output value becomes x multiplied by λ. If the input value is less than or equal to 0, there is a function that tends to 0, which is the output y, when x is zero. Essentially when x less than 0, the alpha is multiplied with the exponential of the x-value minus the alpha value, and then we multiply by the lambda value.

The most important aspect of SELU function is that it is self-normalizing. To be direct, first the mean is subtracted, and then divided by the standard deviation. So the components of the network (weights, biases and activations) will have a mean of zero and a standard deviation of one after the normalization. This is the output of the SELU activation function.

There is an advantage of using the mean of zero and the standard deviation of one. Let us assume that the initialization function LeCun_Normal initializes the parameters of the network as a normal distribution. In the case of SELU, the network will be normalized entirely. Essentially, when multiplying or adding components of such a network, the network will be still considered as a normal distribution. In turn, the whole network and its output in the last layer is also normalized. The graph of a normal distribution with a mean of 0 and a standard deviation of 1.

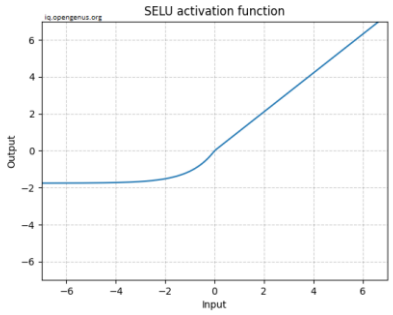

Fig. 6.2.: Scaled Exponential Linear Unit (SELU)

The output of the SELU is normalized, which could be called the internal normalization, hence the fact that all the outputs are with a mean of zero and a standard deviation of one. This is different from external normalization, where the batch normalization and other methods are used. The internal normalization actually happens with SELU because the variance decreases when the input is less than 0 and increases when the input is greater than 0. The standard deviation is the square root of variance, and hence the result is 1.

ISSN- 2394-5125 VOL 09, ISSUE 04, 2022

SELUs allow to construct a mapping g with properties that lead to Self Normalizing Neural Networks (SNN)[7]. SNN cannot be derived by another activation functions other than SELU, which has

- negative and positive values for controlling the mean
- saturation regions(derivatives approaching zero) to dampen the variance of its too large in the lower layer
- a slope larger than 1 to increase the variance, if it is too small, in the lower layer
- a continuous curve.

The differentiated function for the SELU activation is given as:

$$\text{SELU'}(x) = \lambda \begin{cases} 1 & \text{if } x > 0 \\ ae^{x} & \text{if } x \leq 0 \end{cases}$$

(i.e), if x>0, then the output will be y. But if x less than 0 then the alpha value is simply multiplied by the exponential operation on x. The SELU function can be used as the activation unit in the output layer of the Convolutional Neural Network.

4. Conclusion:

The CNNs have proved to differentiate emotions better. These networks have been trained on a predefined set of emotions, be it the 6 basic emotion categories or 3 continuous dimensions, or a random set of 26 emotion categories or compound emotions. The future scope of this paper would be to recommend the use of Rectified Linear Units (ReLU) and the Scaled Exponential Linear Unit (SELU) with the CNN to recognize the 26 emotion categories under the 3 continuous dimensions. The latter being a new area of research, will open new doors for various other CNN models with greater efficiency. This can be further taken up to recognizing emotions without any prior information on the data sets and hence succeed in *detecting depression in individuals and diagnosis of development disorders in children*.

5. References:

[1] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. Journal of personality and social psychology, 17(2):124, 1971

[2] J. Guo et al. Dominant and Complementary Emotion Recognition From Still Images of Faces, Special Section On Visual Surveillance And Biometrics: Practices, Challenges, And Possibilities, Digital Object Identifier 10.1109/ACCESS.2018.2831927, IEEE Access.

[3] I. Lüsi et al., ‘‘Joint challenge on dominant and complementary emotion recognition using micro emotion features and head-pose estimation: Databases,’’ in Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG), May/Jun. 2017, pp. 809–813.

[4] S. Du, Y. Tao, and A. M. Martinez, ‘‘Compound facial expressions of emotion,’’ Proc. Nat. Acad. Sci. USA, vol. 111, no. 15, pp. E1454–E1462, 2014.

[5] A. Martinez and S. Du, ‘‘A model of the perception of facial expressions of emotion by humans: Research overview and perspectives,’’ J. Mach. Learn. Res., vol. 13, no. 1, pp. 1589–1608, Jan. 2012.

ISSN- 2394-5125 VOL 09, ISSUE 04, 2022

[6] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, ‘‘EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 5562–5570.

[7] S. Li, W. Deng, and J. Du, ‘‘Reliable crowdsourcing and deep localitypreserving learning for expression recognition in the wild,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 2584–2593

[8] J. Guo et al., ‘‘Multi-modality network with visual and geometrical information for micro emotion recognition,’’ in Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG), May/Jun. 2017, pp. 814–819

[9] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, ‘‘A discriminative feature learning approach for deep face recognition,’’ in Proc. Eur. Conf. Comput. Vis., 2016, pp. 499–515.

[10] Emotion Recognition in Context Ronak Kosti∗ , Jose M. Alvarez† , Adria Recasens‡ , Agata Lapedriza∗ Universitat Oberta de Catalunya∗ Data61 / CSIRO† Massachusetts Institute of Technology

[11] A. Mehrabian. Framework for a comprehensive description and measurement of emotional states. Genetic, social, and general psychology monographs, 1995

[12] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollar, and ´ C. L. Zitnick. Microsoft COCO: common objects in context. CoRR, abs/1405.0312, 2014

[13] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Semantic understanding of scenes through ade20k dataset. 2016

[14] L. F. Barrett, B. Mesquita, and M. Gendron. Context in emotion perception. Current Directions in Psychological Science, 20(5):286–290, 2011.

[15] J. Alvarez and L. Petersson. Decomposeme: Simplifying convnets for end-to-end learning. CoRR, abs/1606.05426, 2016

[16] https://iq.opengenus.org/scaled-exponential-linear-unit/