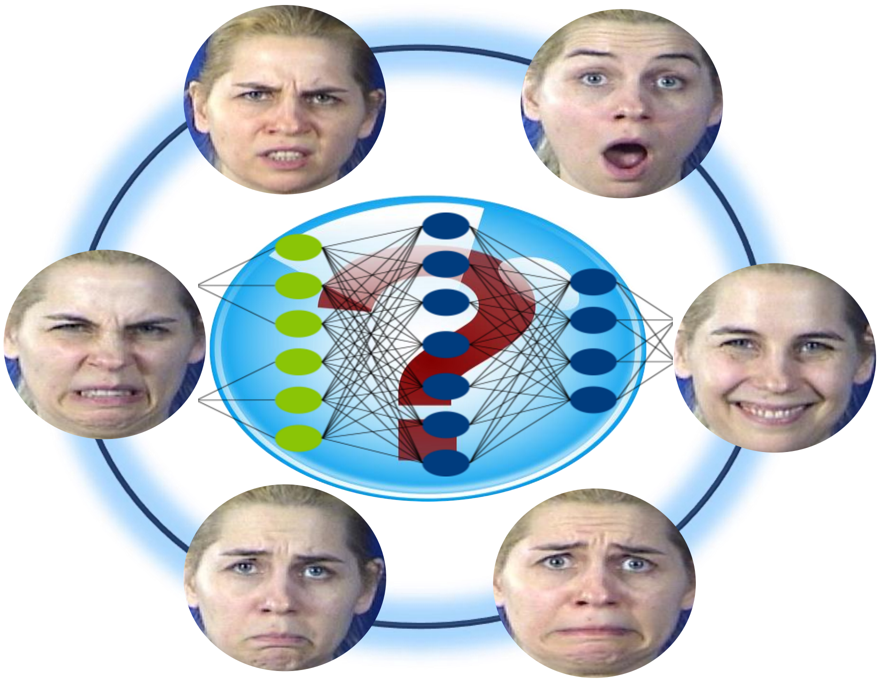
EXPERTNet: Exigent Feature Preservative Network for Facial Expression Recognition

Accepted in 11th ICVGIP_2018 for short oral presentation held in IIIT Hyderabad

By: Monu Verma



Short Falls in existing Networks

EXPERTNet Architecture

Properties of EXPERTNet

Qualitative Analysis

Quantitative Analysis

Conclusion

Short Falls in Existing Networks

- Conventional networks: VGG-16, VGG-19 and ResNet follows deep dense sequential structure. Linearly connected Conv layers may drop some salient feature due to recurrence of cross-correlation, which has an important role to define an expression class. Thus, it degrades the performance of the network
- The GoogleNet utilizes inception layer, which process all different sized filter's responses to learn the best weights when training the network and automatically select the more useful features. Thus, increases computational cost of the network.
- Smaller sized filters 3x3 and 5x5 may loose abstract edge variations, which also plays a significant role in facial expression recognition.
- Conventional CNN-based networks are uses max polling to down sample the input image. Pooling layer extracts the maximum response features by the performing max operation over embedded filters. Thus, max pooling layer also neglects the micro-variation of the facial images.
- Existing networks have large computational cost as they uses large number of learning parameters.

EXPERTNet

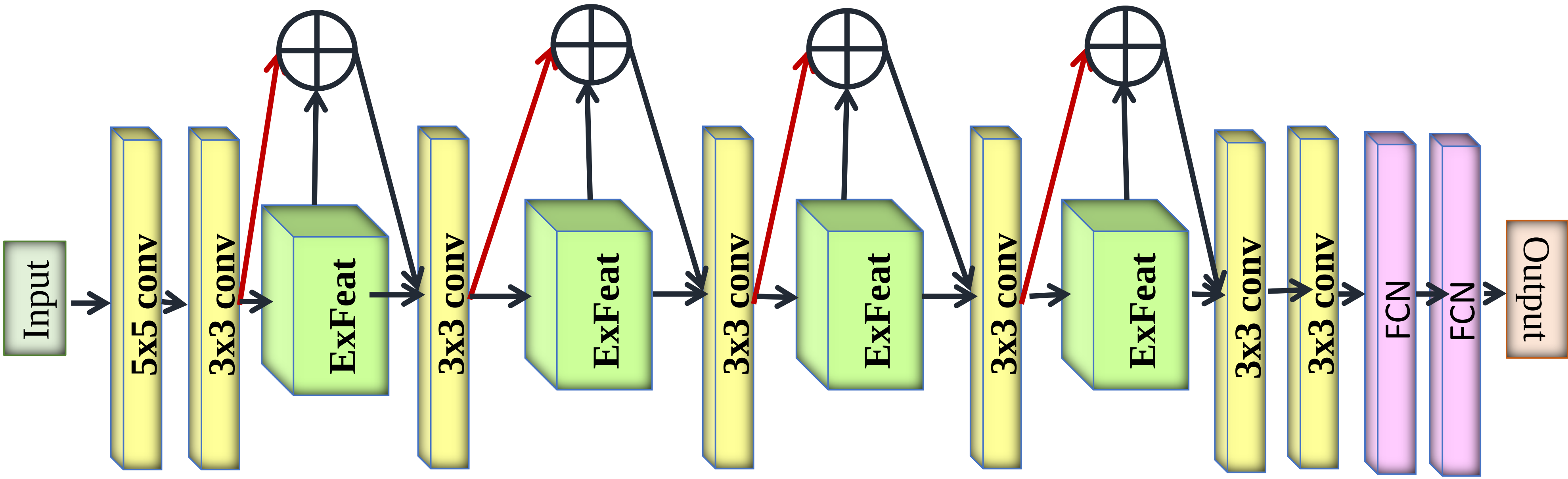
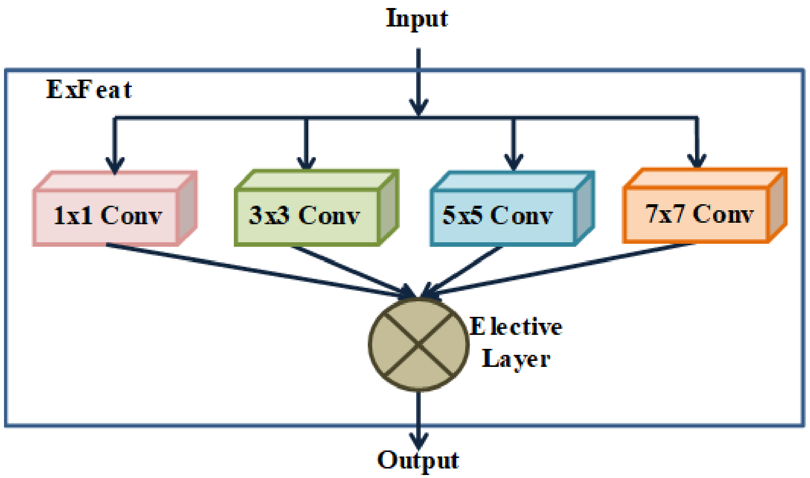
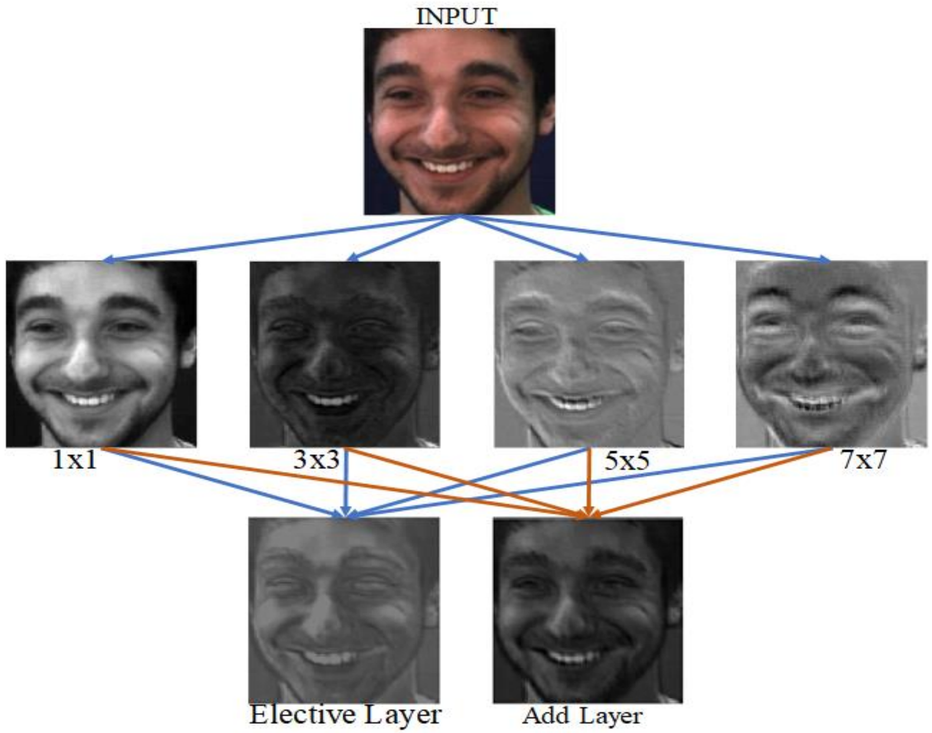


Fig. 1. EXPERTNet Architecture

EXPERTNet-ExFeat Block

ExFeat block, mainly comprises of elective layer, that extracts the pertinent features from both micro and high-level feature responses generated by different sized filters at convolution layer. Elective layer also improves performance of the network by reducing the learning parameters of the hidden layers.





$$R_{\in} = \frac{1}{2}\left((\max(R_n) + (\min(R_n)\right)$$

$$D_n = \left| R_{\in} - R_n \right|$$

$$R_E = R_{\in} + \min(D_n)$$

$$n = \{1, 2, 3, 4\}$$

Mathematical Formulation

The mathematical representation of the network is formulated as:

$$R = R^{5,5,32}_{Conv/1}\left\{I(p,q)\right\} \quad (1)$$

$$R^t_{Feat} = \left(R^{3,3,d}_{Conv/2}(R)\right) \quad (2)$$

$$R^t_{Ex} = ExFeat\left\{R^{2z-1,2z-1,d}_{Conv/1}(R^n_{Feat})\right\}^4_{z=1} \quad (3)$$

where, $t=\{1,2,3,4\}$, $d=\{32,64,96,128\}$

ExFeat is calculated by using Eq. (4-6)

$$ExFeat(R_n) = \phi(R_n) + \min(\chi(R_n)) \quad (4)$$

$$\chi(R_n) = \left|\phi(R_n) - R_n\right| \quad (5)$$

$$\phi(R_n) = \frac{1}{2}((\max(R_n) + (\min(R_n)) \quad (6)$$

Then, final neurons are calculated by using Eq. (7-8)

$$R^t_{Add} = R^t_{Feat} + R^t_{Ex} \quad (7)$$

$$R_{Final} = Fc^7\left(Fc^{1024}\left(Fc^{512}\left(R^{3,3,256}_{Conv/2}\left(R^{3,3,184}_{Conv/2}(R_{Add})\right)\right)\right)\right) \quad (8)$$

Convolution Layer

$$R^{u,v,N}_{Conv/S}\left\{I(p,q)\right\} = \sum_{m=-v/2}^{v/2}\sum_{n=-u/2}^{u/2} f_k(m,n) \otimes I(\alpha-m,\beta-n) \quad (9)$$

$$\begin{cases}\alpha = (S\times p-(S-1)) \\ \beta = (S\times q-(S-1))\end{cases} \quad (10)$$

Vision Intelligence Lab

EXPERTNet- Detailed Architecture

TABLE I
EXPERTNet Detailed Configuration

Layers		Filter	Output	# Param
Input Image		-	128x128x3	-
Conv 1		5x5	128x128x32	2K
Conv 2		3x3	64x64x32	9K
ExFeat 1	Conv 3.1	1x1	64x64x32	86K
	Conv 3.2	3x3		
	Conv 3.3	5x5		
	Conv 3.4	7x7		
Addition 1		-	64x64x32	-
Conv 4		3x3	32x32x64	18K
ExFeat 2	Conv 4.1	1x1	32x32x64	342K
	Conv 4.2	3x3		
	Conv 4.3	5x5		
	Conv 4.4	7x7		
Addition 2		-	32x32x64	-
Conv 5		3x3	16x16x96	55K
ExFeat 3	Conv 6.1	1x1	16x16x96	773K
	Conv 6.2	3x3		
	Conv 6.3	5x5		
	Conv 6.4	7x7		
Addition 3		-	16x16x96	-
Conv 7		3x3	8x8x128	111K
ExFeat 4	Conv 8.1	1x1	8x8x128	1M
	Conv 8.2	3x3		
	Conv 8.3	5x5		
	Conv 8.4	7x7		
Addition 4		-	8x8x128	-
Conv 9		3x3	4x4x184	212K
Conv 10		3x3	2 2 256	424K
Fully Connected 1		-	1x1x512	525K
Fully Connected 2		-	1x1x1024	525K

Properties of EXPERTNet

- ExFeat blocks, incorporates different sized filters 1x1,3x3,5x5 and 7x7. Combination of filters allows extracting both micro and high-level edge features.

- ExFeat blocks contain Elective layer to preserve only exigent feature responses and processed to next layer, instead of all feature responses like Inception layer.

- Additive layer utilizes to combine, response feature maps of prior and current convolution layer to enrich the generated feature responses like ResNet.

- EXPERTNet included convolution layer with stride 2, which decrease the size of input with minimum information loss as shown in Figure. 2.

Fig 2: Input image and response images generated by applying (a) Conv with stride 1 (b) pooling with stride 2 and (c) Conv with stride 2.

Qualitative Analysis

Fig. 3 Visualization of response feature maps for a) neutral b) anger c) disgust d) fear e) happy f) sad and g) surprise expression, capture at elective layer over MMI dataset.

Qualitative Analysis

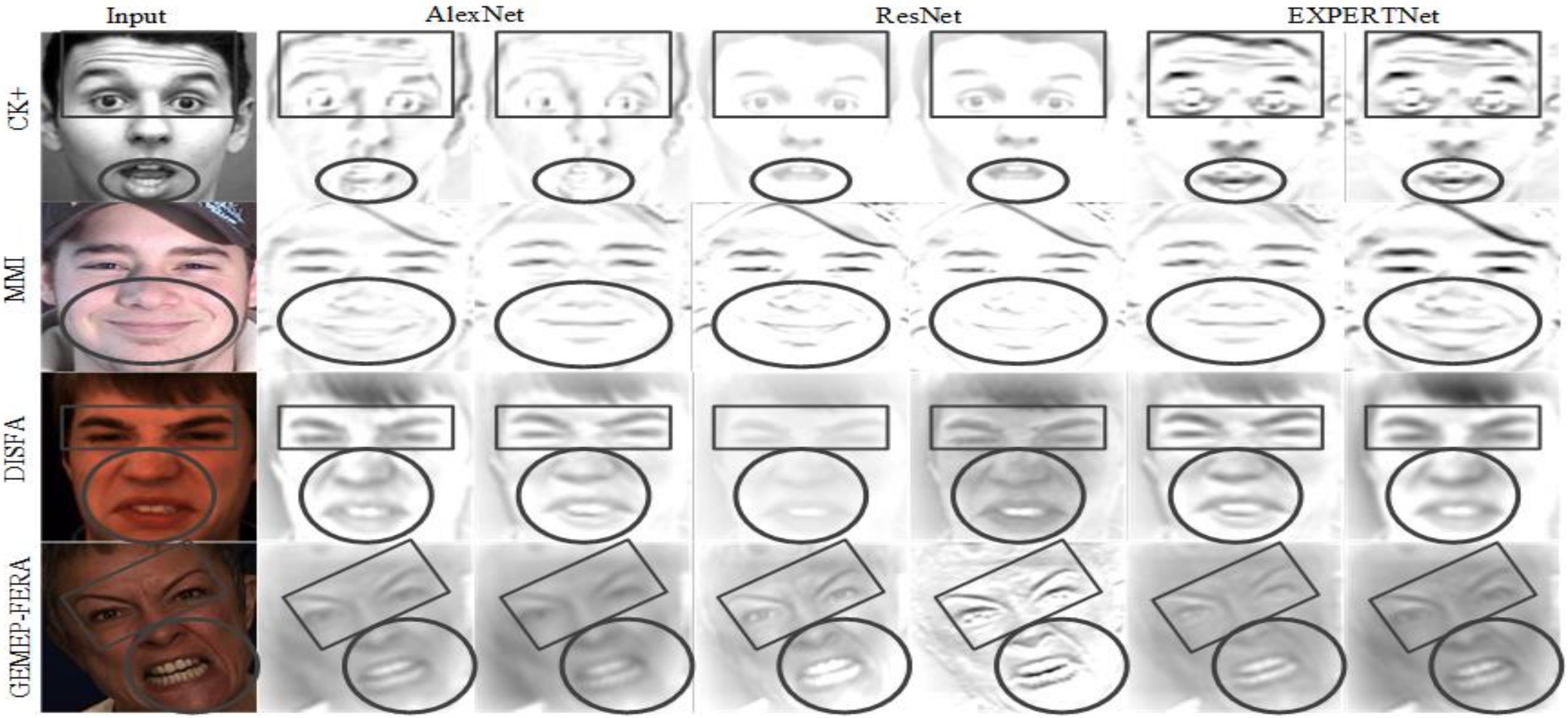


Fig. 4 Visual comparison of existing models and EXPERTNet over different expression of four datasets a) CK+: Surprise b) MMI: Happy c) DISFA: Disgust and d) GEMEP-FERA: Anger.

Quantitative Analysis

TABLE II
Recognition Accuracy Comparison on CK+ Dataset

Method	Accuracy Rate (%)	
	6-class	7-class
LBP [7]	93.5	89.0
Two-Phase [8]	88.2	79.5
LDP [9]	96.2	92.9
LDN [10]	94.8	91.7
LDTexP [11]	95.3	91.9
LDTerP [12]	95.7	91.5
VGG-Net 16 [14]	96.7	95.2
VGG-Net 19 [14]	97.2	81.2
ResNet [16]	94.0	91.8
EXPERTNet	**99.1**	**98.8**

TABLE III
Recognition Accuracy Comparison on MMI Dataset

Method	Accuracy Rate (%)	
	6-class	7-class
LBP [7]	76.5	81.7
Two-Phase [8]	75.4	82.0
LDP [9]	80.5	84.0
LDN [10]	80.5	83.0
LDTexP [11]	83.4	86.0
LDTerP [12]	80.6	80.0
VGG-Net 16 [14]	83.9	89.2
VGG-Net 19 [14]	81.6	83.9
ResNet [16]	71.2	83.9
EXPERTNet	**99.1**	**98.0**

Quantitative Analysis

TABLE IV
Recognition Accuracy Comparison on DISFA Dataset

Method	Accuracy Rate (%)	
	6-class	7-class
LBP [7]	91.8	92.7
Two-Phase [8]	91.0	92.9
LDP [9]	91.5	94.1
LDN [10]	90.7	93.0
LDTexP [11]	92.2	93.8
VGG-Net 16 [14]	89.2	83.9
VGG-Net 19 [14]	83.9	88.3
ResNet [16]	83.9	71.2
EXPERTNet	**95.3**	**95.5**

TABLE V
Recognition Accuracy Comparison on GEMEP-FERA Dataset

Method	Accuracy Rate (%)	
	5-class	6-class
LBP [7]	92.2	87.8
Two-Phase [8]	88.6	85.0
LDP [9]	94.0	90.0
LDN [10]	93.4	91.0
LDTexP [11]	94.0	91.8
VGG-Net 16 [13]	85.1	90.7
VGG-Net 19 [14]	91.8	89.3
ResNet [16]	78.4	78.7
EXPERTNet	**94.4**	**92.9**

Computational Complexity

TABLE VI
Complexity analysis

Network	# Layers	# Parameters
VGG-16	16	138M
VGG-19	19	144M
GoogleNet	22	4M
ResNet	34	11M
EXPERTNet	13	4M

Vision Intelligence Lab

Conclusion

- The EXPERTNet extracts only pertinent features and neglect others by using exigent feature (ExFeat) block, mainly comprises of elective layer.
- ExFeat block contains different sized filters as 1x1, 3x3, 5x5 and 7x7 to capture both local and abstracted features. Thereby, the response feature maps can easily extract the edge variations of facial appearance.
- EXPERTNet combines the former layer response with currently processed layer responses to secure more feature information. Thus, resultant feature maps have capability to define disparities between different expression classes.
- Experimental results have proved effectiveness of the proposed network over four datasets: CK+, MMI, DISFA and GEMEP-FERA.

References

[1] P. Ekman, (1993). Facial expression and emotion. *American Psychologist*, *48*(4), 384.

[2] P. Ekman and W.V. Friesen, (1977). Facial action coding system.

[3] W. V. Friesen and P. Ekman, (1983). EMFACS-7: Emotional facial action coding system. *Unpublished manuscript, University of California at San Francisco*, *2*(36), 1.

[4] I. A. Essa and A. P. Pentland, (1997). Coding, analysis, interpretation, and recognition of facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, *19*(7), 757-763.

[5] I. Kotsia and I. Pitas, (2007). Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE transactions on image processing*, *16*(1), 172-187.

[6] D. Gabor, (1946). Theory of communication. Part 1: The analysis of information. *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, *93*(26), 429-441.

[7] C. Shan, S. Gong and P.W. McOwan (2009). Facial expression recognition based on local binary patterns: A comprehensive study. *Image and vision Computing*, *27*(6), 803-816.

[8] C. C. Lai and C. H. Ko, (2014). Facial expression recognition based on two-stage features extraction. *Optik-International Journal for Light and Electron Optics*, *125*(22), 6678-6680.

[9] T. Jabid, M. H. Kabir and O. Chae, (2010). Robust facial expression recognition based on local directional pattern. *ETRI journal*, *32*(5), 784-794.

[10] A. R. Rivera, J. R. Castillo and O. O. Chae, (2013). Local directional number pattern for face analysis: Face and expression recognition. *IEEE transactions on image processing*, *22*(5), 1740-1752.

[11] A. R. Rivera, J. R. Castillo and O. Chae (2015). Local directional texture pattern image descriptor. *Pattern Recognition Letters*, *51*, 94-100.

[12] B. Ryu, A. R. Rivera, J. Kim and O. Chae, (2017). Local directional ternary pattern for facial expression recognition. *IEEE Transactions on Image Processing*, *26*(12), 6006-6018.

[13] A. Krizhevsky, I. Sutskever and G. E Hinton, (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097-1105.

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

[15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818-2826.

[16] K. He, X. Zhang, S. Ren and J. Sun (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770-778.

References

[17] A. Mollahosseini, D. Chan and M. H. Mahoor, (2016). Going deeper in facial expression recognition using deep neural networks. In *Applications of Computer Vision (WACV),* 1-10.

[18] B. Hasani and M. H. Mahoor, (2017). Spatio-temporal facial expression recognition using convolutional neural networks and conditional random fields. In *12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017),* 790-795.

[19] H. Jung, S. Lee, S. Park, I. Lee, C. Ahn and J. Kim, (2015). Deep temporal appearance-geometry network for facial expression recognition. *arXiv preprint arXiv:1503.01532*.

[20] P. Khorrami, T. Paine, and T. Huang, (2015). Do deep neural networks learn facial action units when doing expression recognition?. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 19-27.

[21] H. Ding, S. K. Zhou and R. Chellappa, (2017). Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In *12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017),* 118-126.

[22] Y. Kim, B. Yoo, Y. Kwak, C. Choi and J. Kim, (2017). Deep generative-contrastive networks for facial expression recognition. *arXiv preprint arXiv:1703.07140*.

[23] P. Burkert, F. Trier, M. Z. Afzal, A. Dengel and M. Liwicki, (2015). Dexpression: Deep convolutional neural network for expression recognition. *arXiv preprint arXiv:1509.05371*.

[24] K. Zhang, Y. Huang, Y. Du and L. Wang, (2017). Facial expression recognition based on deep evolutional spatial-temporal networks. *IEEE Transactions on Image Processing*, *26*(9), 4193-4203.

[25] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, (2010). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* 94-101).

[26] M. Pantic, M. Valstar, R. Rademaker and L. Maat, (2005,). Web-based database for facial expression analysis. In *IEEE international conference on multimedia and Expo*, 5.

[27] M. Valstar and M. Pantic, (2010). 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*, 65.

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

[29] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic and K. Scherer, (2011). The first facial expression recognition and analysis challenge. In *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011),* 921-926.

[30] P.Viola, and M. J. Jones, (2004). Robust real-time face detection. *International journal of computer vision*, *57*(2), 137-154.

Thank You

QUERIES ?