G2-VER: Geometry Guided Model Ensemble for Video-based Facial Expression Recognition

Tanguy Albrici[1], Mandana Fasounaki[2], Saleh Bagher Salimi[1], Guillaume Vray[1],
Behzad Bozorgtabar[1], Hazım Kemal Ekenel[2], Jean-Philippe Thiran[1]
[1]École Polytechnique Fédérale de Lausanne, Switzerland, [2]Istanbul Technical University, Istanbul, Turkey

Abstract— **This paper addresses the problem of automatic facial expression recognition in videos, where the goal is to predict discrete emotion labels best describing the emotions expressed in short video clips. Building on a pre-trained convolutional neural network (CNN) model dedicated to analyzing the video frames and LSTM network designed to process the trajectories of the facial landmarks, this paper investigates several novel directions. First of all, improved face descriptors based on 2D CNNs and facial landmarks are proposed. Second, the paper investigates fusion methods of the features temporally, including a novel hierarchical recurrent neural network combining facial landmark trajectories over time. In addition, we propose a modification to state-of-the-art expression recognition architectures to adapt them to video processing in a simple way. In both ensemble approaches, the temporal information is integrated. Comparative experiments on publicly available video-based facial expression recognition datasets verified that the proposed framework outperforms state-of-the-art methods. Moreover, we introduce a near-infrared video dataset containing facial expressions from subjects driving their cars, which are recorded in real world conditions.**

I. INTRODUCTION

Facial expression recognition (FER) in facial images and videos plays an important role in numerous applications in human-computer interaction, health care, and advanced driver-assistance systems (ADAS). Although this task is widely studied and much progress has been made, it still remains a challenging problem, due to the complexity and variability of facial expressions. So far, most applications of machine learning in FER have been on still frames, to name a few [15], [32], [9].

Many of the so-called traditional local features, such as HOG, SIFT, LBP used in image processing can be extended to be applied to video processing as well. These include, among others, 3D HOG [14], 3D SIFT [29], and LBP-TOP [36]. For example, Liu et al. [19] introduced the UMM (Universal Manifold Model), where a video is expressed as a spatial-temporal manifold (STM) based on a combination of local SIFT and HOG features. This method achieves reasonable performances on benchmark dataset, with 95.10% accuracy on the CK+ dataset [22]. Most tests for video-based FER have been performed on the CK+ dataset [22], the Oulu-CASIA [35] dataset, and the MMI dataset [25], all of which contain RGB videos. In this paper, we introduce a new near-infrared video dataset containing facial expression from subjects driving their cars recorded in real world conditions.

Many state-of-the-art techniques for video-based FER incorporate recurrent neural network (RNN) to model temporal cues from faces. As an example, Kahou et al. [4] developed combination of the CNN-RNN architecture, where the spatial features for each frame of a video are fed to a RNN to extract the spatio-temporal features. In addition, in various works [12], [34] facial landmark trajectory has been shown to be an effective low-level feature for FER. For example, Jung et al. [12] designed a Deep Temporal Geometry Network (DTGN), which has shown that face landmark trajectory fed to a shallow fully-connected network can lead to state-of-the-art results that outperform hand-crafted features.

Motivated by these observations, we proposed two ensemble models that are built on a pre-trained CNN model dedicated to analyze video frames. First of all, improved face descriptors based on 2D CNNs and facial landmarks are proposed (early-fusion model). Second, this paper investigates fusion methods of the features temporally, including a novel hierarchical recurrent neural network combining facial landmark trajectories over time. In addition, we propose modifications to state-of-the-art facial expression recognition architectures to adapt them to video processing in a simple way (late fusion model). Comparative experiments on publicly available video-based FER datasets verified that the proposed framework outperforms state-of-the-art methods.

II. Related Work

Early studies in FER were mostly focused on emotion recognition in static images. However, new methods have been proposed for recognizing facial expressions in videos. A review of the FER approaches are given in the following subsections.

A. *Conventional FER Approaches*

Many of traditional image features used for FER are either geometry-based features, such as facial landmarks or appearance-based features, such as Gabor filters [17] and Local Binary Patterns [30]. Geometry-based features represent the shape of the face and its components, whereas appearance-based features describe the texture of the face. For example, [37], [11], [3] used local binary pattern (LBP) descriptor as image features. These extracted features are then fed into classifiers such as support vector machine (SVM) [8], [33] or k-nearest-neighbor algorithm (KNN) [28] to classify the input images into discrete emotion categories. To formulate the spatio-temporal evolution of human facial

978-1-7281-0089-0/19/$31.00 ©2019 IEEE

expressions, different ensemble models are used to model the variability in morphological and contextual factors. The method presented in [6] uses multi-class AdaBoost and SVM for facial expression recognition in image sequences.

B. FER Using Deep Convolutional Neural Networks

Unlike traditional approaches, deep learning methods such as convolutional neural networks demonstrated significantly high performances on different computer vision tasks, such as image classification [15], [32], [9]. Therefore, in most recent studies [21], [31], CNN models have been used for FER in static images as well. In [13] Jung et al. utilized a method to capture the dynamical variations of expressions with two deep neural networks (DTAGN). In order to model the temporal trajectories of facial landmarks, hybrid deep learning approaches, combining different architectures of CNNs and RNNs are proposed. Liu et al. in [18] proposed expressionlet-based spatio-temporal mainfold descriptor for dynamic facial expression recognition. Zhang et al. [34] proposed a system combining a part-based hierarchical bidirectional recurrent neural network (PHRNN) to extract dynamic geometry information and a multi-signal convolutional neural network (MSCNN) to boost the performance of facial expression recognition from consecutive frames. Ebrahimi et al. [5], also used a combination of CNNs and RNNs for FER in videos. Long short-term memory (LSTM) cell, which is a refinement to the basic recurrent neural network architecture is used in [7] for facial expression recognition. Majumder et al. in [24] take benefit of data fusion technique. In this study, fusion of geometric features and LBP features is created using autoencoders and Kohonen self organizing map (SOM)-based classifier [23], and high accuracy is achieved on CK+ and MMI [26] databases.

III. DATASETS

In this study, we have employed three face datasets for different purposes and experiments. In the following subsections these datasets are described.

A. MUG Dataset

The MUG Facial Expression Database [1] consists of 938 short videos of 52 subjects performing facial expressions. Each subject has between 11 and 26 videos and each emotion has between 52 and 162 videos. This dataset contains 7 basic emotions (anger, disgust, fear, happiness, sadness, surprise, neutral), and each video starts and ends at the neutral state.

B. CK+ Dataset

The Extended Cohn-Kanade Expression Database (CK+) [22] consists of 327 short videos of 118 subjects performing facial expressions. Each subject has between 1 and 6 videos and each emotion has between 19 and 82 videos. This dataset contains 7 basic emotions (anger, disgust, fear, happiness, sadness, surprise, and contempt), and the peak expression is at the end of each video. This is one of the most popular dataset for FER and it is thus very important in order to compare our methods with the state of the art.

C. Driver Face Dataset

The driver face dataset consists of sequences of near-infrared videos of subjects driving their cars in real world conditions. This dataset includes three Use Cases (UC), each one representing a specific vehicle category, where several automated systems interact with the user. They cover different types of vehicles including use case A (truck), use case B (electric car) and use case C (conventional car). There are 3 subjects for each use case, making a total of 9 subjects. Since this is a real-world dataset, it brings a few additional challenges compared to the other two public datasets. The main challenge is the class imbalance, since the neutral emotion is more represented than any other, as seen in Table I. Another challenge is that in some cases, the face is partly occluded by a microphone, or sensors on the subject's face, as seen in Fig. 1.

Neutral	4000 (59%)
Positive	1192 (18%)
Negative	1575 (23%)
TOTAL	6767 (100%)

TABLE I

CLASS DISTRIBUTION FOR THE DRIVER FACE DATASET.

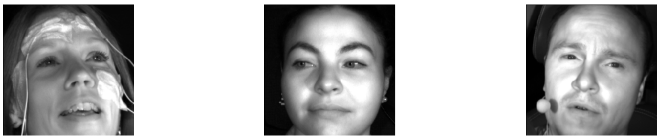

Fig. 1. Sample frames from three different use cases of the driver face dataset.

IV. EARLY FUSION MODEL

A. Overall Architecture

In our first model, we explore the early fusion of CNN-based features and handcrafted features. The motivation behind this approach is that in most cases, the amount of data required for CNN-based methods are much more important than for conventional FER methods. However, the available data for facial expression recognition, such as CK+, is limited and the problem is that CNN-based methods tend to overfit in such cases. On the other hand, handcrafted feature extraction methods such as the Scale Invariant Feature Transform (SIFT) do not require extensive datasets to generalize. Therefore, we propose a hybrid approach by combining CNN and SIFT features to get the best of both worlds. The former are extracted using a deep convolutional neural network pre-trained on a dataset of facial images, while the latter are obtained by the SIFT features around the facial landmarks. Since SIFT features are extracted from facial landmarks, we can benefit from the geometry-based information in the data. For each frame in the image sequence that serves as input to the model, these two types of features are computed, concatenated, and then passed through an LSTM network, followed by fully connected layers that output the final emotion prediction. The overall architecture of this model is shown in Fig. 2.

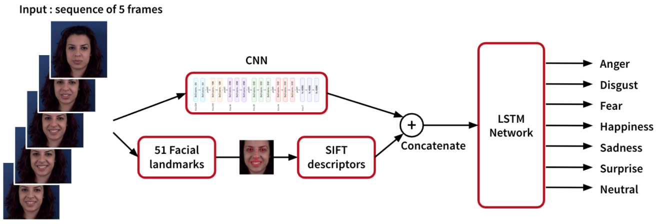

Fig. 2. Overall architecture of the Early Fusion model. Five consecutive frames demonstrating one emotion, are on one hand fed into the CNN model to extract features. On the other hand, face landmarks of the faces in all of the frames are detected and SIFT features are extracted of the landmarks. The concatenation of these two features are then used to training the LSTM network.

B. *Customized CNNs*

To choose an optimal architecture that is suitable for the extraction of the CNN-based features, we evaluated the performance of multiple popular deep CNNs on the Real-world Affective Faces (RAF) database [16], which consists of very diverse facial images downloaded from the Internet. For each of the pre-trained CNNs, we fine-tuned the whole network using training images of the RAF database and reported results on the test set of this dataset. We observed that, as shown in Table II, the best performance was obtained using VGG-Face [27] and SqueezeNet [10], therefore, we selected these two architectures to extract the CNN-based features. Even though we obtain a slightly lower accuracy with SqueezeNet, its main advantage is that its inference speed is much faster than VGG-Face, since it is a very compact network. These two backbone CNNs are also used in the second model (Late Fusion), described in Section V.

TABLE II
PERFORMANCE OF DIFFERENT CNNS ON THE RAF DATASET

CNN	Accuracy
VGG-Face	**86%**
SqueezeNet	**83%**
DenseNet 121	82%
Inception V3	82%
Xception	81%
VGG19	83%
NAS Net	83%

C. *Facial Landmarks and SIFT*

The handcrafted features are obtained by first detecting 51 facial landmarks on the image, and then computing the SIFT descriptors of each of these 51 landmarks.

V. LATE FUSION MODEL

A. *Overall Architecture*

In the second model, our approach is to explore the late fusion of CNN-based and handcrafted features. Here, two separate neural networks are trained and later combined by averaging their score vector. Since the main motivation is to use temporal information of the facial expression structure, we combined the prediction scored of a customized CNN called Temporal CNN (TCNN) and a Part-based Hierarchical Recurrent Neural Network (PHRNN). The overall architecture of Late Fusion model is shown in Fig. 3.

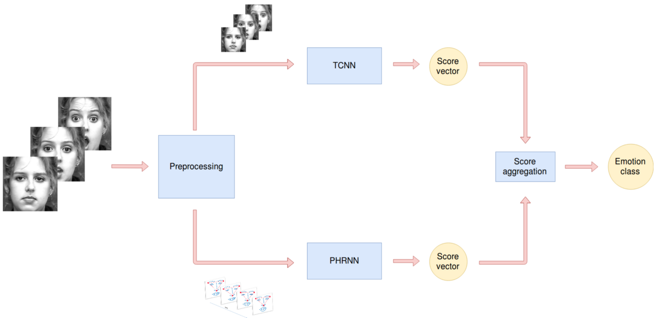

Fig. 3. Overall architecture of the Late Fusion model. Figure from [2]

B. *Temporal CNN (TCNN)*

The temporal CNN is simply a customized version of our base CNN (either VGG-Face or SqueezeNet) using pre-trained weights, modified to be able to take multiple frames as input. The only modification to the base CNN architecture is therefore at the first layer, where one convolution (followed by a non-linearity) is applied to each of the input frames. Then the results of these convolutions are aggregated by averaging and from this point on, the standard architecture of the base CNN is used, as shown in Fig. 4.

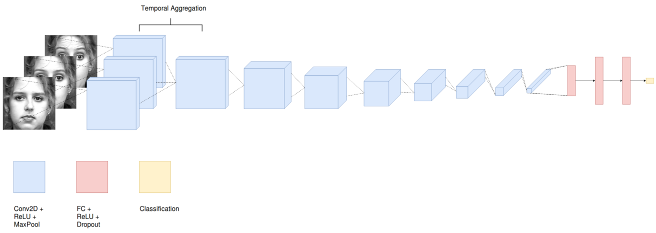

Fig. 4. Architecture of the TCNN model. Figure from [2]

C. *Part-Based Hierarchical RNN (PHRNN)*

This model uses the facial landmarks of each frame as input (either as normalized landmark coordinates or as SIFT descriptors). The normalization is made by subtracting the coordinates of the tip of the nose, and then dividing the resulting X and Y coordinates by their respective standard deviation. The particularity of this model is that clusters of facial landmarks are hierarchically grouped in order to better model spatial dependencies. Temporal dependencies are modeled by using several Bidirectional-RNNs and a Bidirectional-LSTM at the very end.

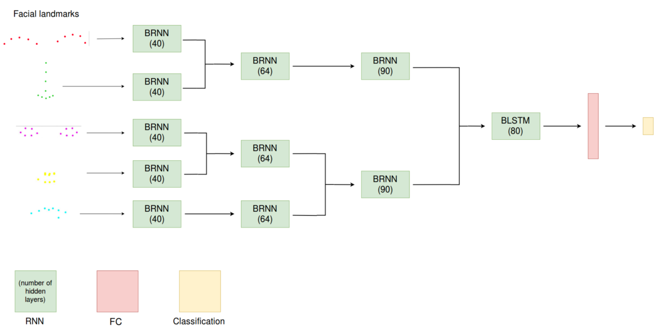

Fig. 5. Architecture of the PHRNN model. Figure from [2]

VI. Results

A. MUG Dataset

1) Evaluation Setup: For the MUG dataset, we take as input of our models 5 frames centered around the peak expression of the image sequence. Higher and lower number of frames were tested and 5 was chosen as the best one. The performance of the model is evaluated using 5-fold cross-validation. The 5 splits are taken so that they have roughly the same number of samples (about 188), while maximizing the subject independence. For the TCNN of the late fusion model, we use VGG-Face as the basis model.

2) Results: We can observe in the table below that the early fusion and late fusion models show a very similar performance. The confusion matrices are shown in the Appendix (section VIII-A)

TABLE III
Results on the MUG Dataset

	Accuracy	Precision	Recall	F1-Score
Early Fusion Model	95.9%	95.7%	93.7%	94.5%
Late Fusion Model	95.5%	94.7%	94.2%	94.4%
TCNN	92.9%	92.0%	91.6%	91.7%
PHRNN	90.8%	90.1%	89.5%	89.9%

B. CK+ Dataset

1) Evaluation Setup: For comparison purposes, we use the same setup as the work of Liu et al. [20], which is the state-of-the-art on the CK+ dataset. We take as input of our models the last 3 frames of the image sequence. The CK+ dataset is split into 8 subsets in a strict subject independent manner, and 8-fold cross-validation is used. For each fold, 6 subsets are used for training and the 2 remaining are used for validation. For the TCNN of the late fusion model, we use VGG-Face as the basis model.

2) Results: We can observe in the table below that the late fusion model performs better on the CK+ dataset. With an accuracy of 97.4%, we outperform the state-of-the-art [20]. The confusion matrices are shown in the Appendix (section VIII-B)

TABLE IV
Results on the CK+ Dataset

	Accuracy	Precision	Recall	F1-Score
Early Fusion Model	96.3%	94.5%	93.2%	93.8%
Late Fusion Model	97.4%	96.7%	96.4%	96.4%
TCNN	94.2%	93.5%	93.0%	93.7%
PHRNN	92.5%	91.8%	91.8%	91.7%
State-of-the-art [20]	97.1%	N/A	N/A	N/A

C. Driver Face Dataset

1) Evaluation Setup: The input consists of 5 consecutive frames labelled with the same emotion. To evaluate the performance of the models, we use a leave-one-driver-out cross-validation. The used data contains 9 different drivers, and the results below (metrics and confusion matrices) take all of the 9 test sets into account. Only the late fusion model was tested on this dataset. Both the VGG-Face and SqueezeNet versions of the TCNN are evaluated in distinct late fusion models that we respectively call *PHRNN + VGG-TCNN* and *PHRNN + SqueezeNet-TCNN*.

2) Results: We can observe in the table below that the two Late Fusion models have a quite similar performance, but *PHRNN + VGG-TCNN* is slightly better. The confusion matrices of these two models are shown in the Appendix (section VIII-C)

TABLE V
Results on the driver face dataset

	Accuracy	Precision	Recall	F1-Score
PHRNN + VGG-TCNN	80.2%	75.5%	79.8%	77.3%
PHRNN + SqueezeNet-TCNN	79.8%	75.2%	76.5%	75.8%
PHRNN	75.9%	70.7%	72.8%	71.5%
VGG-TCNN	74.6%	69.2%	74.8%	71.4%
SqueezeNet-TCNN	66.6%	58.9%	62.0%	60.0%

VII. Conclusion

We presented two ensemble models building on pre-trained CNNs dedicated to automatic emotion recognition on video frames. The first one, early fusion, uses improved face descriptors based on 2D CNNs and facial landmarks. The second one, late fusion, includes a novel hierarchical recurrent neural network combining facial landmark trajectories over time, and presents a way to adapt frame-wise expression recognition architectures to videos in a simple way. Comparative experiments on publicly available video-based facial expression recognition datasets verified that the proposed framework outperforms state-of-the-art methods. Finally, we introduced a near-infrared video dataset containing facial expression from subjects driving their cars recorded in real world conditions.

Acknowledgments

This work is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 688900 (ADAS&ME project - http://www.adasandme.com).

VIII. APPENDIX

A. MUG Dataset

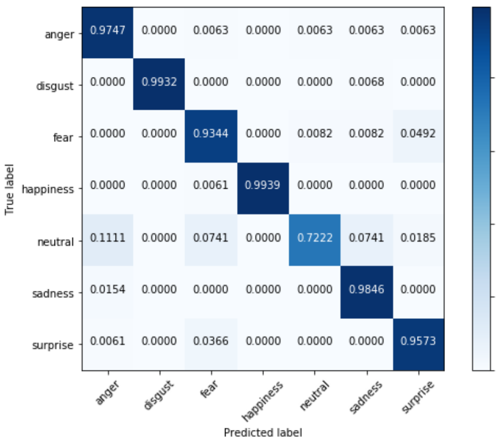

Fig. 6. Confusion matrix for the Early Fusion model on the MUG dataset.

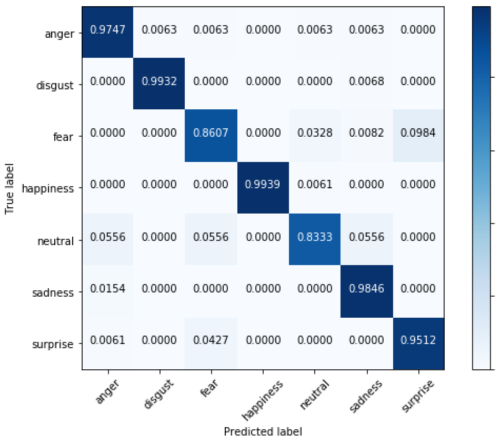

Fig. 7. Confusion matrix for the Late Fusion model on the MUG dataset.

B. CK+ Dataset

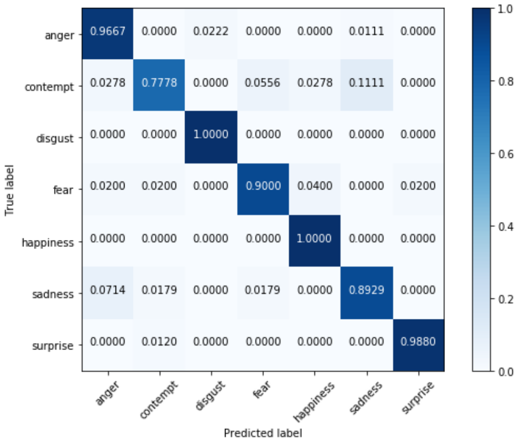

Fig. 8. Confusion matrix for the Early Fusion model on the CK+ dataset.

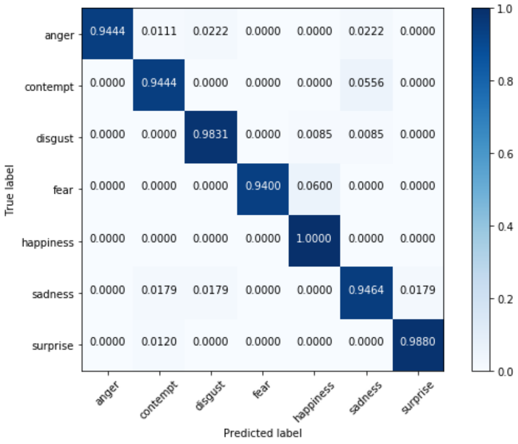

Fig. 9. Confusion matrix for the Late Fusion model on the CK+ dataset.

C. Driver Face Dataset

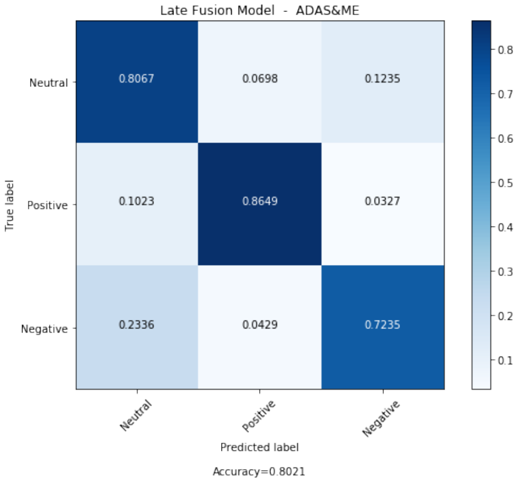

Fig. 10. Confusion matrix for the Late Fusion (PHRNN + VGG-TCNN) model on the driver face dataset.

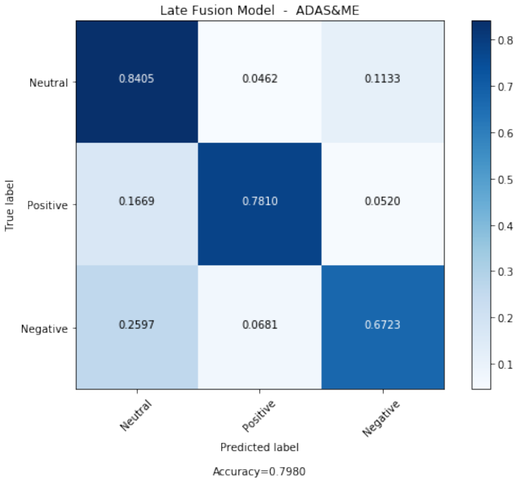

Fig. 11. Confusion matrix for the Late Fusion (PHRNN + SqueezeNet-TCNN) model on the driver face dataset.

References

[1] N. Aifanti, C. Papachristou, and A. Delopoulos. The mug facial expression database. In *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*, pages 1–4, April 2010.

[2] K. Bacsa, J.-P. Thiran, and A. Sugimoto. Video-based emotion recognition using deep learning techniques. *Master Thesis*, 2018.

[3] M. J. Cossetin, J. C. Nievola, and A. L. Koerich. Facial expression recognition using a pairwise feature selection and classification approach. In *Neural Networks (IJCNN), 2016 International Joint Conference on*, pages 5149–5155. IEEE, 2016.

[4] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal. Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 467–474. ACM, 2015.

[5] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal. Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 467–474, New York, NY, USA, 2015. ACM.

[6] D. Ghimire and J. Lee. Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines. *Sensors*, 13(6):7714–7734, 2013.

[7] A. Graves, J. Schmidhuber, C. Mayer, M. Wimmer, and B. Radig. Facial expression recognition with recurrent neural networks. 2008.

[8] S. Happy and A. Routray. Automatic facial expression recognition using features of salient facial patches. *IEEE transactions on Affective Computing*, 6(1):1–12, 2015.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[10] F. N. Iandola, M. W. Moskewicz, K. Ashraf, S. Han, W. J. Dally, and K. Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <1mb model size. *CoRR*, abs/1602.07360, 2016.

[11] Y. Ji and K. Idrissi. Automatic facial expression recognition based on spatiotemporal descriptors. *Pattern Recognition Letters*, 33(10):1373–1380, 2012.

[12] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2983–2991, 2015.

[13] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[14] A. Klaser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC 2008-19th British Machine Vision Conference*, pages 275–1. British Machine Vision Association, 2008.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, USA, 2012. Curran Associates Inc.

[16] S. Li, W. Deng, and J. Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2593. IEEE, 2017.

[17] C. Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Transactions on Image Processing*, 11(4):467–476, April 2002.

[18] M. Liu, S. Shan, R. Wang, and X. Chen. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1749–1756, 2014.

[19] M. Liu, S. Shan, R. Wang, and X. Chen. Learning expressionlets via universal manifold model for dynamic facial expression recognition. *IEEE Transactions on Image Processing*, 25(12):5920–5932, 2016.

[20] X. Liu, B. V. K. V. Kumar, J. You, and P. Jia. Adaptive deep metric learning for identity-aware facial expression recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 522–531, July 2017.

[21] A. T. Lopes, E. de Aguiar, A. F. D. Souza, and T. Oliveira-Santos. Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order. *Pattern Recognition*, 61:610 – 628, 2017.

[22] 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*, pages 94–101, June 2010.

[23] A. Majumder, L. Behera, and V. K. Subramanian. Emotion recognition from geometric facial features using self-organizing map. *Pattern Recognition*, 47(3):1282 – 1293, 2014. Handwriting Recognition and other PR Applications.

[24] A. Majumder, L. Behera, and V. K. Subramanian. Automatic facial expression recognition system using deep network-based data fusion. *IEEE Transactions on Cybernetics*, 48(1):103–114, Jan 2018.

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

[26] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *2005 IEEE International Conference on Multimedia and Expo*, pages 5 pp.–, July 2005.

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

[28] A. Poursaberi, H. A. Noubari, M. Gavrilova, and S. N. Yanushkevich. Gauss–laguerre wavelet textural feature fusion with geometrical information for facial expression identification. *EURASIP Journal on Image and Video Processing*, 2012(1):17, Sep 2012.

[29] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th ACM international conference on Multimedia*, pages 357–360. ACM, 2007.

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

[31] K. Shan, J. Guo, W. You, D. Lu, and R. Bie. Automatic facial expression recognition based on a deep convolutional-neural-network structure. In *Software Engineering Research, Management and Applications (SERA), 2017 IEEE 15th International Conference on*, pages 123–128. IEEE, 2017.

[32] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[33] H.-H. Tsai and Y.-C. Chang. Facial expression recognition using a combination of multiple facial features and support vector machine. *Soft Computing*, 22(13):4389–4405, 2018.

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

[35] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. PietikäInen. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9):607–619, 2011.

[36] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):915–928, 2007.

[37] G. Zhao and M. Pietikinen. Boosted multi-resolution spatiotemporal descriptors for facial expression recognition. *Pattern Recognition Letters*, 30(12):1117 – 1127, 2009. Image/video-based Pattern Analysis and HCI Applications.