Interpretable Emotion Classification Using Temporal Convolutional Models

Manasi Gund*
Rochester Institute of Technology
Rochester, New York 14623
Email: mg9546@rit.edu

Abhiram Ravi Bharadwaj*
Rochester Institute of Technology
Rochester, New York 14623
Email: raviabhiram@rit.edu

Ifeoma Nwogu
Rochester Institute of Technology
Rochester, New York 14623
Email: ionvcs@rit.edu

***Abstract*—As with many problems solved by deep neural networks, existing solutions rarely explain, precisely, the important factors responsible for the predictions made by the model. This work looks to investigate how different spatial regions and landmark points change in position over time, to better explain the underlying factors responsible for various facial emotion expressions. By pinpointing the specific regions or points responsible for the classification of a particular facial expression, we gain better insight into the dynamics of the face when displaying that emotion. To accomplish this, we examine two spatiotemporal representations of moving faces, while expressing different emotions. The representations are then presented to a convolutional neural network for emotion classification. Class activation maps are used in highlighting the regions of interest and the results are qualitatively compared with the well known facial action units, using the facial action coding system. The model was originally trained and tested on the CK+ dataset for emotion classification, and then generalized to the SAMM dataset. In so doing, we successfully present an interpretable technique for understanding the dynamics that occur during convolutional-based prediction tasks on sequences of face data.**

I. Introduction

In interacting with faces, although a significant amount of information can be conveyed by static faces, it is possible that some information can be better transmitted via dynamic faces. This is also a more realistic depiction of how we interact and interpret information from faces. Some useful applications of modeling dynamic instead of static faces include analyzing spontaneous facial expressions [1], automatically detecting pain via facial dynamics [2], recognizing delighted versus frustrated smiles [3], etc.

In this work, we explore how the temporal patterns on the face can provide additional insight in to the underlying dynamics of the facial behavior when expressing emotion. To accomplish this, we train 2 different temporal image representations using the CK+ dataset. The CK+ dataset [4] provides image sequences extracted from videos along with the represented emotion class. A total of 7 emotions were classified. The model trained and tested on CK+ was applied on the SAMM dataset [5], to demonstrate that it was did not overfit on the CK+ dataset. SAMM also contained 7 emotions recorded over multiple sequences.

The first temporal face image representation is referred to as the stacked convolutional network (SCN), while the other which uses extracted facial landmarks, is referred to as the temporal convolutional network (TCN). The common network used in training both representations are inspired from the VGG-16 [6] architecture. When tested on the CK+ dataset, both these models give accuracies of over 95%, although only the results obtained using facial landmarks match up closer with the state-of-the-art results on the benchmark datasets. To better understand what landmarks the models "believe" are important for prediction, class activation maps [7] were computed on test samples, after successfully training the models. Figure 1 shows one such instance depicting the highly activated regions obtained from the TCN model. The image shows the the importance of the various facial landmarks, in classifying the surprise emotion.

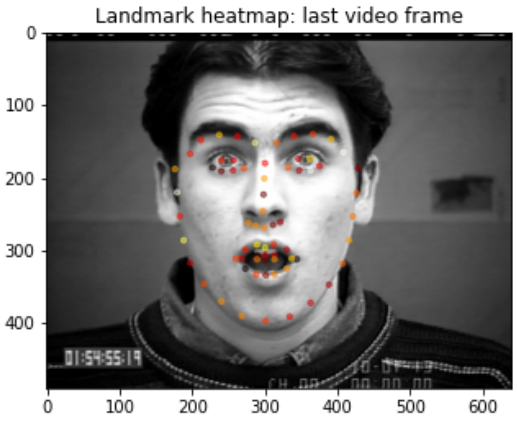


Fig. 1: (a) An example of a subject portraying the emotion of surprise with the activations for the temporal convolutional netowork on the image shown. The activations vary from white $\rightarrow$ yellow $\rightarrow$ red, where the more red the landmark, the more activated it is. The model correctly predicted that the expression being demonstrated by the subject was *surprise*.

Specifically, in this work we are interested in classifying sequences of face images by the emotions they display and using class activation maps (CAMs), we can explore the facial dynamics that best explain the classification.

*Equal contribution

II. Background

Although a great deal of work has been done in the field of facial expression recognition, not as much work has been done using the dynamism of the faces for this work. This is probably due to the fact that one can predict facial expressions to a large extent using single images. In this work though, we are interested in evaluating facial expression recognition in moving faces.

There are several different approaches to dynamic facial expression recognition, varying from classifications using simple multi-layer perceptrons [8] to complicated deep neural networks. Many of such solutions involve splitting the videos into frames and passing them to a convolutional neural network [9] one frame at a time. Others, pass the extracted frames to a recurrent model, after extracting features from a convolutional neural network [10]–[13]. Other works detect facial action units and process them to detect emotions [14], [15]. The work of Fuzail Khan [8] utilises a simple multi-layer perceptron with landmarks to classify the emotions in images, however, the landmarks considered are just the four fiducial features - the eyebrows, the eyes, the nose and the lips. While videos can be broken into frames, this loses the temporal aspect of the video. The entire temporal history of the landmarks can be quite useful for many face-based prediction tasks including emotion classification and techniques that use only frame based processing lose this history. In his work, Shivam Gupta [16] used a support vector machine to classify emotions based on landmarks. While classifying the emotion, the landmarks are not used individually but rather the center of gravity of the collection is computed. Then the location of the center of gravity is used to classify the image into the appropriate emotion. This approach although unique runs the risk of losing out on the information contained in the individual locations of the landmarks.

Minaee et al. [17] used stacked images with a CNN to show the regions in the image that were most activated during classification.The TCN model utilised by He et al. [18] used the location of facial landmarks to detect emotions. He et al. extracted the landmarks over video frames and then computed the differences in landmark locations. Since facial landmarks are nothing but points in the video frame, they simply subtract the location of one landmark from it's location in the next frame and used this as the input to training the network. While a useful approach, it has the possibility of aggregating the location of landmarks, thus neglecting the history of the landmarks.

Though not directly related to emotion recognition in moving faces, other techniques for detecting facial features include the work by [19] who use independent components analysis (ICA) is to learn the appearance and shape of the facial features; work by [20] that present a comprehensive survey on facial feature points detection; another work on feature detection by [21]

The proposed work involves recognizing the expressed emotion on the face, using moving landmark points. Although the positions of the landmarks on a static image can successfully aid in detecting facial expression, tracking those landmarks over a period of time (in dynamic faces) has not being investigated in as much detail. Some of the more common techniques such as the use of LSTM and LSTM combined with CNNs have been investigated, hence, in this paper, we explore some novel approaches to recognize facial expressions in moving faces. Although these techniques are not novel in pattern recognition in general, they have not been explored in general facial analysis. In this work, we explore a couple of interpretable methods for

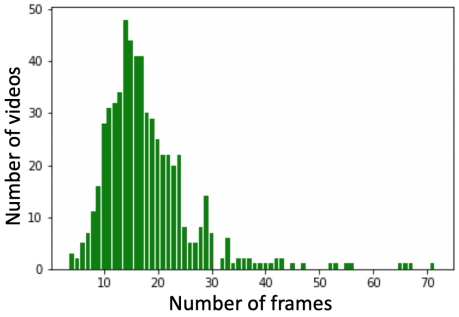


Fig. 2: The varying number of frames among all the videos available in the CK+ dataset.

III. Data

A. CK+ dataset

The CK+ dataset is primarily used for training both, the SCN and the TCN. The CK+ dataset provides four sets of information - the frames containing the face image, the corresponding landmarks file for the frames, the FACS encodings for the frames and the emotion label. The dataset was created by asking subjects to portray various emotions. The subjects were required to go from a neutral expression to the extreme of their given emotion in a span of one minute. In total, 593 such videos were recorded. The videos were split into frames which were made available as the dataset. The number of frames for each video was not constant and this variation is shown in Figure 2, which depicts the number of videos for different frame counts. All of the 593 did not however have labels and only 327 emotion sequences were available for training and testing; the remaining 266 sequences did not readily translate into the standard emotion prototypes, and could therefore not be readily labeled. The CK+ dataset provided image sequences for seven emotions: *anger, contempt, disgust, fear, happy, sad and surprise*.

B. SAMM dataset

The SAMM dataset consists of videos for micro and macro emotions, although for this work, we only use the macro expression labels in our extended testing. Similar to the CK+ dataset, the SAMM dataset provides image sequences of subjects which are extracted from videos. The difference being that these videos were recorded to portray spontaneous

Sr. No.	Class	Count
1	Anger	39
2	Contempt	12
3	Disgust	46
4	Fear	20
5	Happy	62
6	Sad	21
7	Surprise	69

TABLE I: Table showing the different number of video samples available for each emotion after filtering.

expressions, hence each sequence did not restrict the subject to a single emotion. The dataset however does provide annotations which indicate the exact frame where the emotion started (onset), the frame the expression peaked (apex), the frame where the emotion went back to normal (offset) and what the emotion being depicted is. The dataset also provides the action units that were observed to have ben elicited by the subjects. In total, the SAMM dataset provided 343 macro emotion videos with labels.

IV. Method

A. Pre-processing

The information provided in both the datasets were not readily usable out of the box. One of the major issues with the CK+ dataset was that the number of frames for the videos are not all the same. This would not bode well with any model being developed since the models require a fixed input size. Based on the distribution of frames per video, we standarised all videos to 20 frames. It was also noted that while downsampling would not cause much of an issue, upsampling might introduce redundancies into the model. Hence all videos with less than 10 frames were dropped from consideration. Table I gives a final count of the number of videos that were eventually used for training.

The CK+ dataset and the SAMM dataset videos were provided with the same classes of emotions. For this research however, we decided that 'contempt' would not be considered because there was only a small number of instances in the databases. Neither of the datasets provided data points for a neutral emotion. Although all videos went from neutral to some other expression, none of them stayed neutral through the duration of the video. Since the video frames were provided in the datasets it was easy to extract the neutral frames to create a new class.

To create neutral data points for training, 50 videos were selected at random from the CK+ dataset and 5 videos (image sequence for expressions) were selected from the SAMM dataset. The first frame from each one of those were chosen and replicated 20 times. These replications were stored in a folder of its own, similar to all the other video frames given by CK+ dataset. Their corresponding labels were also created in files that adhered the folder structure of CK+ dataset.

The SAMM dataset was constructed using a camera which recorded 200 frames per second, resulting in a large number of frames for small video segments. To avoid any additional computational expenses, for each of the emotions required, 20 frames were subsampled based on the apex and onset. For each emotion, 5 videos were selected at random and 20 frames were extracted.

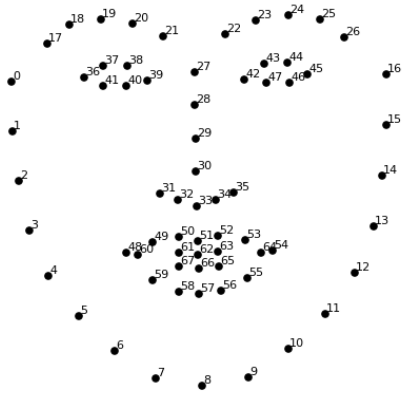


Fig. 3: A skeleton of a face showing the 68 facial landmark locations as provided by OpenFace. Image from [22].

The SCN takes image frames as its input and both datasets used in this study provide gray-scale images. Hence the input representation to this stacked model did not require much pre-processing. The TCN however takes as input facial landmarks. Although the CK+ dataset came with pre-generated landmarks, because the source-code for their landmarks generation was not provide), we regenerated our own version using the toolkit, OpenFace [22], to maintain consistency between the two datasets. Figure 3 shows the 68 landmarks points on the face as provided by OpenFace. An additional step that was also done here prior to storing the data was to normalize the facial landmark location with respect to the size of the image. This ensured uniformity in the data. For the SAMM dataset, facial landmarks were only generated for the subsampled frames.

Figure 4 shows a schematic diagram of how the input image is transformed and fed into the TCN model.

B. Input representations

1) Stacked Convolutional Model (SCN): The first model is the stacked convolutional model. This model, unlike other video emotion classification models does not use any form of recurrent layers. The input is a set of 20 frames that are stacked together. The image dimension was restricted to 224×224. The work done by Minaee et al. [17] is similar to this model in the sense that they also used stacked images as input and did not use any recurrent model for classification. The underlying convolutional network is inspired by VGG-16. This model consists of four blocks, the first three blocks are similar, having two mini-blocks of convolutional layers followed by a batch

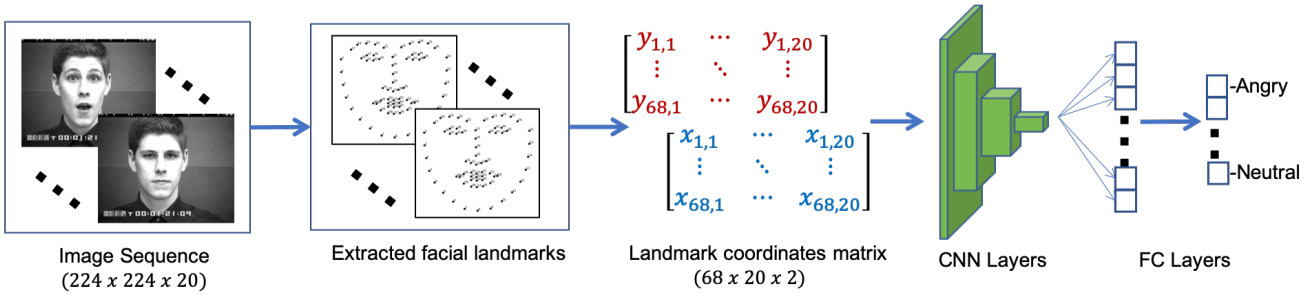


Fig. 4: The temporal convolutional model which primarily takes in as input a metrix containing facial landmarks.

normalization layer. The two mini-blocks are then followed by a max-pooling layer. All activations in these three blocks use ReLU. All convolutional layers in the first, second and third blocks comprise of 64, 128 and 256 filters respectively. All of these filters are $\{3 \times 3\}$. The max pooling filters are sized $\{2 \times 2\}$ and have a stride of 2. The fourth block consists of two fully connected layers. The first one has 256 nodes and again uses ReLU for activation. The second fully connected layer has 7 nodes corresponding to the 7 classes and uses Softmax loss. In total, there are 5,647,943 parameters that are trainable.

2) Temporal Convolution Network (TCN): The TCN looks to capture the temporal aspects of emotion classification without the use of complex recurrent networks. The input to this model, just like with the SCN, is a set of 20 frames with the difference being that it takes only the facial landmark locations instead of the actual images. Since each frame has 68 landmarks, each consisting on an $x-$ and a $y-$ coordinate, the input to the TCN model is of dimension $\{68 \times 20 \times 2\}$. Unlike SCN, the TCN model consists of three blocks. The first block is similar to the first block of the SCN with the difference being that the first convolutional layer has 64 $\{3 \times 3\}$ filters and the second has 128 $\{2 \times 2\}$ filters. The max pooling layer has a pool size of $\{3 \times 3\}$ with a stride of 2. The second block is just one of the mini-block mentioned in the SCN. The final block has two fully connected layer. The first one has 64 hidden units and the second one has 7 hidden units. All the layers that use activations make use of ReLU except for the last one which uses the Softmax loss as well. In total, the TCN model learns on 330,119 parameters.

V. Experiments and Results

A. Classification

The models were successfully implemented using the stochastic gradient descent optimiser, and we found that the best results were obtained when the learning rate was set to 1e-3 and a rate decay of 1e-4. To prevent stagnation during training, we had to set the right momentum to avoid plateauing. For the SCN model, the momentum was set at 0.7 while for the TCN model at 0.3.

To prevent the model from overfitting, 6-fold cross validation was performed where 15% of the CK+ dataset was held out for validation each time. The SCN model yielded an accuracy of **95.67%** while the TCN model yielded **99.57%**.

Figure 5 shows the confusion matrix obtained on testing with the validation set of CK+ dataset. Since the TCN model gives better results and is a more novel approach, the figures for the SCN model are omitted.

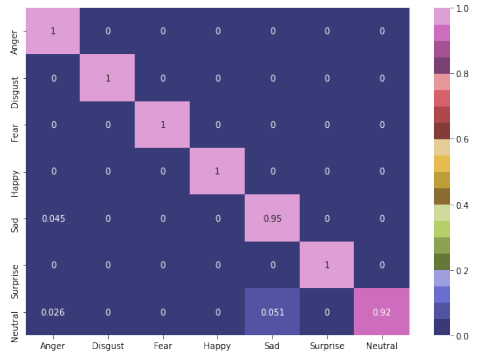


Fig. 5: The confusion matrix obtained when tested on the CK+ dataset with the TCN model.

Table II shows the results obtained by the two models on the CK+ dataset. While there were situations where the SCN model outperformed TCN, the latter is overall the better performing model.

Emotion	Precision		Recall		f1-score	
	TCN	SCN	TCN	SCN	TCN	SCN
Anger	0.95	1.00	1.00	0.92	0.98	0.96
Disgust	1.00	1.00	1.00	0.93	1.00	0.97
Fear	1.00	1.00	1.00	0.90	1.00	0.95
Happy	1.00	0.98	1.00	0.98	1.00	0.98
Sad	0.91	1.00	0.95	0.92	0.93	0.96
Surprise	1.00	0.88	1.00	0.98	1.00	0.93
Neutral	1.00	0.93	0.92	0.98	0.96	0.95

TABLE II: The classification report obtained on CK+ dataset run on both, the stacked convolutional network (SCN) and the temporal convolutional network (TCN).

Model	CK+	SAMM	
	Accuracy(%)	TP	F1-score
FAN	**99.7**	-	-
DeepEmotion	99.3	-	-
DeepConv [23]	92.8	-	-
TimeConvNets	97.9	-	-
Baseline MEGC [24]	-	22	0.06
SCN	95.7	30	0.16
TCN	**99.6**	**41**	**0.33**

TABLE III: Comparison of SCN and TCN with other models for the two datasets used.

Fig. 6: Class activation map for the happy emotion from the stacked convolution network (SCN). Red indicates areas of high activation. No information can be gleaned from this.

Table III shows how well these two models fare with respect to other models reported in the literature. The numbers in bold are the best performers withing statistical significance. It is important to note the simplicity of the TCN model and the small computational cost it incurs by taking in just landmarks instead of entire images. Compared to all other models that use the CK+ dataset, the TCN model proves to be the simplest with best results.At the time of writing this report, there are few other research works published on the SAMM dataset. The only other one which does is the work of He et al. who set the baseline performance for the SAMM dataset. Their work produced 22 true positives for the 343 macro emotion videos, which the two models built in this project overshoot by a significant margin. The TCN model gets more than five times the F1-score that the baseline advocates.

B. Model Interpretation

Obtaining good accuracy from the models was not the end goal of this work. The primary aim of this project was to develop interpretable models for convolution-based prediction tasks on temporal face data. In order to interpret the models, we employed the use of class activation maps, to better understand what spatiotemporal locations in the input images were responsible for the classification results, for both SCN and TCN.

Figure 6 is the result of plotting the class activation map for a highly scoring prediction of happy on an emotive face. As can be observed from the image, the heat map did not provide any insights into the workings of the SCN, on why this test sample scored so highly, being correctly classified as happy.

Figure 7 shows these some activation results from TCN, where the input image is 68×20 representing the 68 landmark points on the face, over 20 frames.

Post-processing had to be done in order to visualise the activations from TCN. Because the resulting activation values were high, they were normalised to lie in the range [0,255]. And since the main focus of interest is in the change in activations over each time frame, the difference between successive frames was computed. Finally, low activations that were not much different from each other (difference in the range of [-75,75]) were dropped to 0 to improve visualisation. The resulting activation maps along with the apex image for the specific emotion are shown in Figure 7 below. To fully interprete the results and compare with emotion-inducing action units, one would obtain the facial locations of the highly activated landmarks from Figure 3 and compare how these to the expected action units related to the emotion (listed in the left captions of Figure 7.

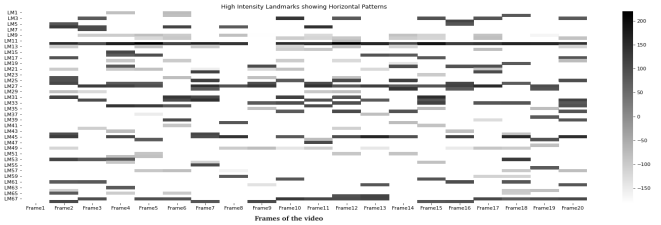
For example, the activation map for anger Figure 7 (g) and (h) demonstrate that several landmarks are involved with the anger emotion and this is depicted on the heat map image. The landmarks also correlate with many of the associated AU regions of interest as listed for the emotion. Similarly, when we look at the neutral expression, the heat map image shows little activity on the face compared to the other images.

The last row is a failure case where the activation map does not correspond to the expected action unit regions involved in sadness. The TCN model correctly predicted the emotions in every case shown here.

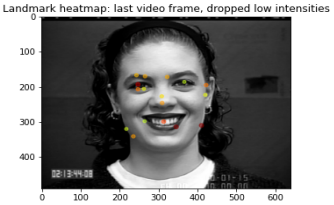
VI. Conclusions

This work presents two convolutional models and analysed them from an interpretability stand point. While traditional convolutional models for facial emotion recognition rely heavily on images as input to the deep networks, this work shows that it might not be necessary as coarse landmark points could still effectively allow for accurate predictions and can also be readily interpreted due to its simplicity.

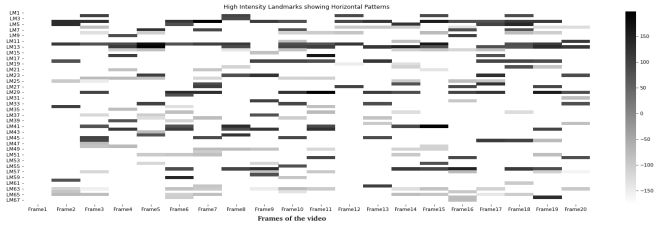
While this work is in no way complete, it serves as a good first step into a new avenue in analysing human facial dynamics. It is possible that coarse facial landmarks such as has been used in this work have been underrated in the literature and there are definitely steps that can be taken to make the system built in this work, especially it interpretability, more robust and yield better results on more generic datasets. A logical next step would be to expand the model to more diverse face-based



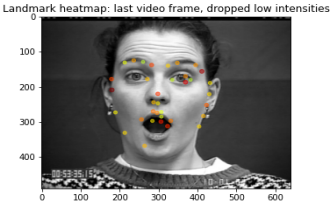
(a) Happiness/joy - AU6 + AU12 - Cheek Raiser, Lip Corner Puller



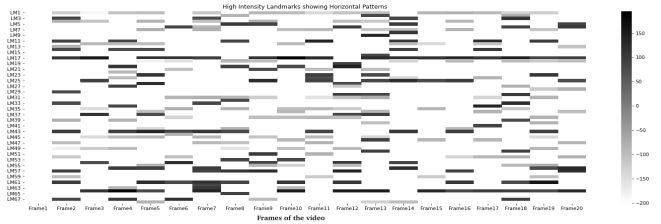
(b) Emotion displayed: Happy



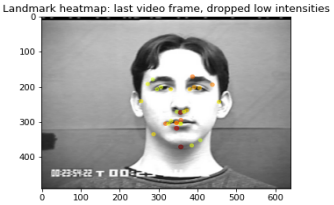
(c) Surprise - AU1 + AU2 + AU5 + AU26 - Inner Brow Raiser, Outer Brow



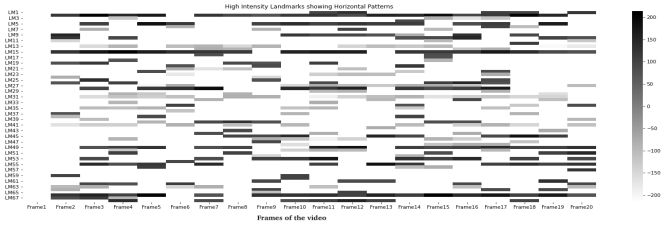
(d) Emotion displayed: Surprise



(e) Neutral - No motion

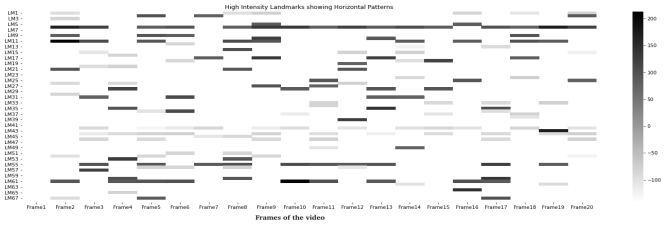


(f) Emotion displayed: Neutral



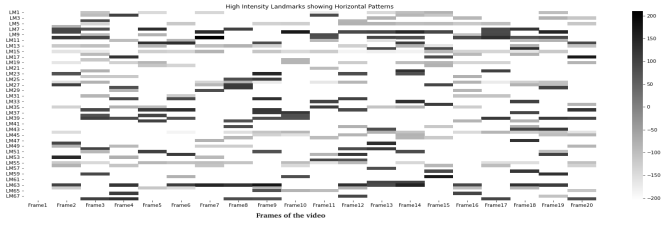
(g) Anger - AU4 + AU5 + AU7 + AU23 - Brow Lowerer, Upper Lid Raiser, Lid Tightener, Lip Tightener

(h) Emotion displayed: Anger

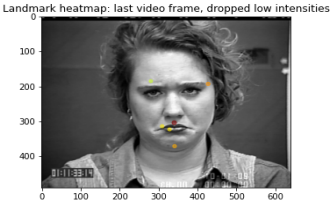


(i) Disgust - AU9 + AU15 + AU16 - Nose Wrinkler, Lip Corner Depressor, Lower Lip Depressor

(j) Emotion displayed: Disgust



(k) Sadness - AU1 + AU4 + AU15 - Inner Brow Raiser, Brow Lowerer, Lip Corner Depressor



(l) Emotion displayed: Sadness

Fig. 7: L: Class Activation Maps for landmarks over the frames of the video. R: CAM on the face.

prediction problems to determine if we can better explain the underlying face-based kinesics such as in the communication of pain, stress, boredom, etc; in the study of neuropsychiatric disorders from the face, and other related problems.

REFERENCES

[1] G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Automatic recognition of facial actions in spontaneous expressions," *Journal of Multimedia*, vol. 1, 09 2006.

[2] M. Aung, S. Kaltwang, B. Romera-Paredes, B. Martinez, A. Singh, M. Cella, M. Valstar, H. Meng, A. Kemp, A. Elkins, N. Tyler, P. Watson, A. Williams, M. Pantic, and N. Bianchi-Berthouze, "The automatic detection of chronic pain-related expression: requirements, challenges and a multimodal dataset," *IEEE Transactions on Affective Computing*, vol. 99, pp. 1–1, 07 2015.

[3] M. E. Hoque, D. J. McDuff, and R. W. Picard, "Exploring temporal patterns in classifying frustrated and delighted smiles," *IEEE Transactions on Affective Computing*, vol. 3, no. 3, pp. 323–334, 2012.

[4] 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*, 2010, pp. 94–101.

[5] C. H. Yap, C. Kendrick, and M. H. Yap, "Samm long videos: A spontaneous facial micro- and macro-expressions dataset," 2019.

[6] S. Liu and W. Deng, "Very deep convolutional neural network based image classification using small training sample size," in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, 2015, pp. 730–734.

[7] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," *CoRR*, vol. abs/1512.04150, 2015. [Online]. Available: http://arxiv.org/abs/1512.04150

[8] F. Khan, "Facial expression recognition using facial landmark detection and feature extraction via neural networks," *CoRR*, vol. abs/1812.04510, 2018. [Online]. Available: http://arxiv.org/abs/1812.04510

[9] J. R. H. Lee and A. Wong, "Timeconvnets: A deep time windowed convolution neural network design for real-time video facial expression recognition," 2020.

[10] M. Sun, S. Hsu, M. Yang, and J. Chien, "Context-aware cascade attention-based RNN for video emotion recognition," *CoRR*, vol. abs/1805.12098, 2018. [Online]. Available: http://arxiv.org/abs/1805.12098

[11] 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*, ser. ICMI '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 467–474. [Online]. Available: https://doi.org/10.1145/2818346.2830596

[12] N. Liu and F. Ren, "Emotion classification using a cnn-lstm-based model for smooth emotional synchronization of the humanoid robot ren-xin," *PLoS ONE*, vol. 14, pp. 53 – 90, 2019.

[13] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using cnn-rnn and c3d hybrid networks," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ser. ICMI '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 445–450. [Online]. Available: https://doi.org/10.1145/2993148.2997632

[14] J. J. Lien, T. Kanade, J. F. Cohn, and Ching-Chung Li, "Automated facial expression recognition based on facs action units," in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, 1998, pp. 390–395.

[15] M. Ghayoumi and A. K. Bansal, "Unifying geometric features and facial action units for improved performance of facial expression analysis," *CoRR*, vol. abs/1606.00822, 2016. [Online]. Available: http://arxiv.org/abs/1606.00822

[16] S. Gupta, "Facial emotion recognition in real-time and static images," in *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, 2018, pp. 553–560.

[17] S. Minaee and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," *CoRR*, vol. abs/1902.01019, 2019. [Online]. Available: http://arxiv.org/abs/1902.01019

[18] Z. He, T. Jin, A. Basu, J. Soraghan, G. Di Caterina, and L. Petropoulakis, "Human emotion recognition in video using subtraction pre-processing," in *Proceedings of the 2019 11th International Conference on Machine Learning and Computing*, ser. ICMLC '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 374–379. [Online]. Available: https://doi.org/10.1145/3318299.3318321

[19] M. Hassaballah and K. Murakami, "Eye and nose fields detection from gray scale facial images," 06 2011.

[20] N. Wang, X. Gao, D. Tao, H. Yang, and X. Li, "Facial feature point detection," *Neurocomput.*, vol. 275, pp. 50–65, Jan. 2018.

[21] M. Hassaballah, S. Bekhet, A. A. M. Rashed, and G. Zhang, "Facial features detection and localization," *Recent Advances in Computer Vision Studies in Computational Intelligence*, pp. 33–59, 2018.

[22] P. Korshunov and S. Marcel, "Speaker inconsistency detection in tampered video," in *2018 26th European Signal Processing Conference (EUSIPCO)*, 09 2018, pp. 2375–2379.

[23] D. Y. Liliana, "Emotion recognition from facial expression using deep convolutional neural network," *Journal of Physics: Conference Series*, vol. 1193, p. 012004, 04 2019.

[24] Y. He, S.-J. Wang, J. Li, and M. H. Yap, "Spotting macro- and micro-expression intervals in long video sequences," 2019.