Human Emotion Recognition in Video Using Subtraction Pre-Processing

Zhihao He
CeSIP, Royal College Building
University of Strathclyde,Glasgow
+44-07712734792
zhihao.he@strath.ac.uk

Tian Jin
CeSIP, Royal College Building
University of Strathclyde, Glasgow
+86-15257220611
tian.jin@strath.ac.uk

Amlan Basu
CeSIP, Royal College Building
University of Strathclyde, Glasgow
+44-7459802138
amlan.basu@strath.ac.uk

John Soraghan
CeSIP, Royal College Building
University of Strathclyde, Glasgow
+44 (0)141 548 2514
j.soraghan@strath.ac.uk

Gaetano Di Caterina
CeSIP, Royal College Building
University of Strathclyde, Glasgow
01415484458
gaetano.di-caterina@strath.ac.uk

Lykourgos Petropoulakis
CeSIP, Royal College Building
University of Strathclyde, Glasgow
+44-7459802138
l.petropoulakis@strath.ac.uk

ABSTRACT

In this paper, we describe a new image pre-processing method, which can show features or important information clearly. Deep learning methods have grown rapidly in the last ten years and have better performance than the traditional machine learning methods in many domains. Deep learning shows its powerful ability particular in difficult multi-classes classification challenges. Video Facial expression recognition is one of the most popular classification topics and will become essential in robotics and auto-motion fields. The new system presented is a combination of new video pre-processing and Convolutional Neural Network (CNN). The new pre-processing method is proposed because we believe individual emotions are dynamic, which means the change of the face is the key feature. RAVDESS is the video set used, to train and test the neural network. From RAVDESS dataset the video songs without audio are taken for focusing on video frames differences. The chosen video set has six different classes of emotions. Each video presents a sentence in a melodious way. Based on the chosen video set, the new system with a new pre-processing method has been designed and trained. Later, the classification result of the new method has been compared with others in which the same dataset for video emotion recognition was used.

CCS Concepts

• **Computing methodologies→ Computer vision.**

Keywords

Classification; Video pre-processing; Images' difference; Emotion Recognition; Neural Networks.

ICMLC '19, February 22–24, 2019, Zhuhai, China

ACM ISBN 978-1-4503-6600-7/19/02…$15.00
DOI: https://doi.org/10.1145/3318299.3318321

1. INTRODUCTION

With the development of GPU, deep learning and back prorogation method became more powerful in classification tasks. Plenty of famous classification models appear (AlexNet, GoogleNet, ResNet) and have good performances. Some models even do better than humans in some challenges. Among all classify objects, distinguish human emotions from video attract our interest. However, human emotions are dynamic and complex, which are not easy to learn, even using deep convolution neural network, which has the best performance in plenty of image databases. The generalization ability of the emotion model is still a serious problem. Previously, Behzad Hasani and Mohammad H. Mahoor [1] created a 3D convolution neural network (CNN) structure using LSTM(Long Short Time Memory) [2] to analyze emotions from videos. Chu et al. [3] proposed a multi-level facial AU detection algorithm that helps to find the features of the face in motion through the time. Graves et al. [4] used a different Recurrent Neural Network (RNN) to analyze temporal information. However, in this work, a new video pre-processing mechanism is developed and tested along with a CNN. The difference of two frames is presented as a result. The backgrounds in the videos are static and static pixels will be black after running the subtract operation. However, the dynamic pixels still show features after subtraction. Other than face detection, face alignment and resizing of the pictures, the pre-processing step also reduces the noise. The video set becomes an image dataset after pre-processing the videos. A well trained CNN is used to train this new image dataset. To have the best classification result, Alex-net, google-net, ResNet structures are used and discussed.

2. RELATED WORK

Facial expression recognition has plenty of methods to achieve this goal. In other words, it is still a classification task. There is a difference between video and image. The video has its temporal relationship between each frame whereas in the image the same does not exist. More and more researchers have paid more attention to the relationship between frames in videos. Our approach is inspired by the traditional video facial recognition shown in the Figure 1

The traditional method has four steps to analyze the video. First, the video is decompressed to the single frame, and then every frame from the previous step reduces the size by using face

detection or landmark detection which only focuses on face information which is shown in figure 1(a) and 1(b). According to the face landmark, facial features are extracted by CNN or other machine learning approach shown in figure 1(c). Finally, the features of each frame are fed into the classifier and the system will show the result of the frame shown in figure 1(d). This method is reasonable, but the drawback of this method is that the whole process does not show the relationship between the frames. The traditional method is just like using an image classification model. Because the traditional method takes more time to extract features and also to calculate the cost of face present in the frame.

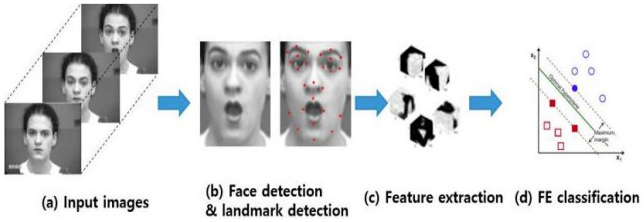

Figure 1. Tradition emotion recognition method [5]

In recent years, researchers presented that the relation of the frame is important and some related works have been done. These works are normally accomplished using CNN to extract image features, and long short-term memory (LSTM) [2] is used to understand temporal sequence features. CNN has state-of-the-art performance in classifying task, and LSTM can analyze the different length of the video sequence. These two powerful algorithms create a system which should be suitable for video analysis. Some works have shown great results using such models. Some of the representative studies using this model are as follows,

Kahou et al. [6] combine an RNN (Recurrent neural network) with a CNN framework. In 2015 Emotion Recognition in the Wild (EmotiW) Challenge [7], the results in the paper show that RNN-CNN system has better performance then deep learning CNN model.

Kim et al. [8] performed the emotion recognition in two parts. The first part is CNN structure to extract spatial features and the second part using the features from the first part to train an LSTM structure to understand the temporal information.

Graves et al. [9] used different LSTM called bidirectional LSTM and unidirectional LSTM to consider the temporal information. The bidirectional LSTM learns video sequence in forward and backward order, which has better performance than single order.

3. OUR APPROACH

From the discussed work in the previous section, we can say that the relation between frames in a video is really important but useless information is still too much. For example, the background of the frames, ears, and hair don't have any contribution to one's emotion recognition. To reduce this, we consider that we add a pre-processing step in videos.

A video is an image sequence; if the camera is fixed, then the background will always be the same. So, if we use two frames, in the video and subtract, the result will show the difference between the frames. This pre-processing method can create a very good image dataset if the frames per second are high enough and the quality of the image is high. Figure 2 shows a skate shoe sample which is created by calculating the difference of the frames in a short video. The background becomes total black with pepper noise. The shape and the details of this shoe are very clearly.

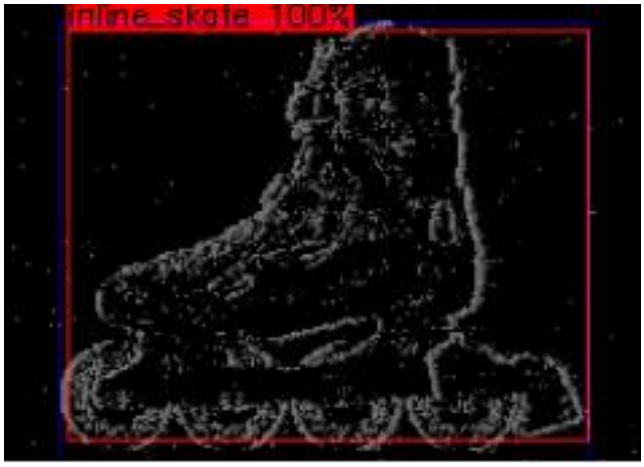

Figure 2. Difference dataset sample with high FPS and high-quality camera [10]

A new image dataset is created by pre-processing video set. The novelty of this paper is to use the difference of the face to classify emotions in the real world and videos. While the whole system still needs CNN to extract images features and other pre-processing operation to reduce the noise and useless information.

For instance, face detection, face alignment.

3.1 Model Structure

This section presents the details about the whole emotions recognition system. Figure 3 shows the sketch of the whole system that includes pre-processing and test progress. The video dataset of RAVDESS [11, 12] is used for pre-processing and training. First, the input video is decomposed into frames and made ready for subtraction. Before subtracting operation, face detection and face alignment technique are used to reduce the background and focus on face part. Subtract operation also needs to set up a gap and a stride value. The gap means the distance between two frames during subtraction. If the gap is too small, then two frames will not have too much difference and mean value of the result after subtraction will be close to zero, which shows that it is losing too many features and cannot be accepted. If the gap is too big, then the relation between frames become weak that means this system's goal is not achieved and won't show a good classification result. The stride in this system is also important, which has almost the same meaning in 1D and 2D convolution. The strides like pooling help in reducing the size of the dataset so that the network does not face the problem of memory size.

After pre-processing, CNN structures are used to do the classification task. AlexNet [13], GoogleNet [14], ResNet [15] structures are used to train and test.

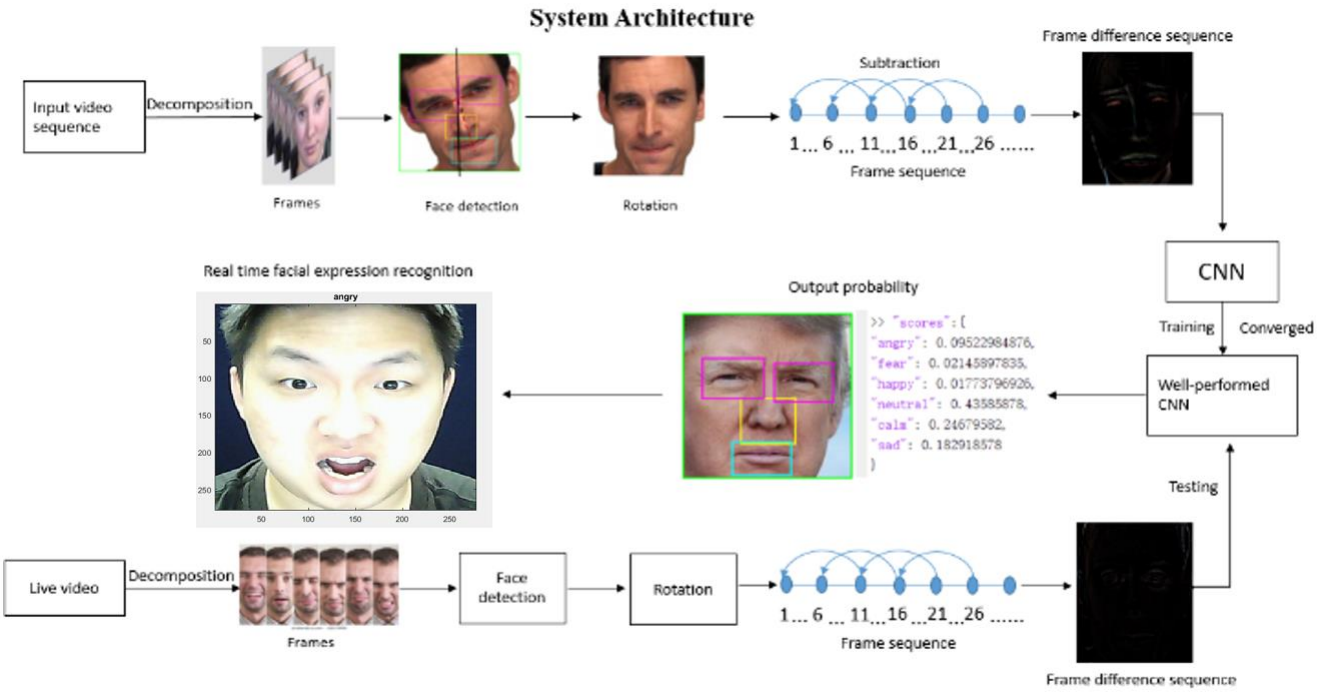

Figure 3. Sketch of new pre-processing system

3.1.1 Face Detection

This system using Haar features detects a face in the picture. Haar-like features method has several steps to detect faces from a picture. First of all, this method needs different filters and set the threshold to decide the face part. Figure 4 shows some haar filters [16], each filter used for one type of edge detection. The second step is face detection algorithm. In this decision algorithm, every small part of the original picture will be split into either not a face or is a face. The part considered as a face will remain and not a face part will be removed. Figure 5 shows a brief decision process. Also, eyes, nose, mouth, can be detected in the same operation. A test of the picture is shown in Figure 6. In Figure 6, nose, eyes and mouth together with the face are detected.

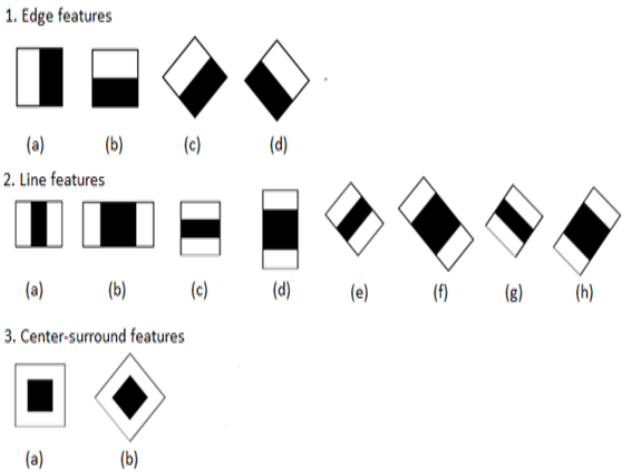

Figure 4. Haar filters sample

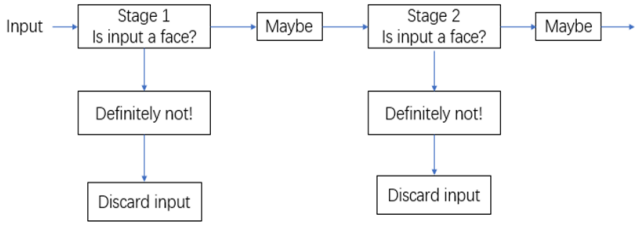

Figure 5. Haar-like decision diagram

In the real test part, we found a great part of the background is cut off. Subtract operation using the images after face detection and face alignment. So the final result also influences the face detection part. If faces are not correct or precious picked from the original images, the CNN cannot find useful features from the difference image, which is generated by the subtraction of frames.

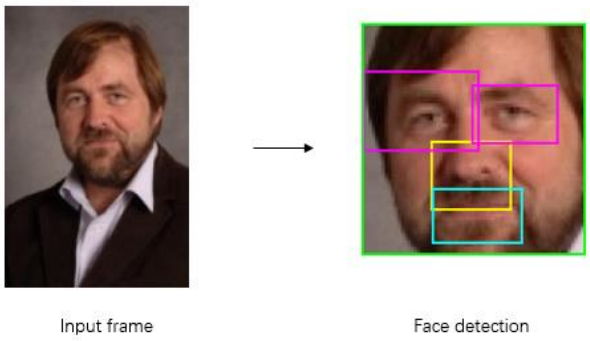

Figure 6. Face detection test

3.1.2 Face Alignment

Face alignment is another pre-processing, which is important and necessary. The volunteers in the video have some gestures when talking or singing, which will cause an error when we subtract the

face. To avoid this, we present an idea to force the face in the image straight by using eyes, nose, and mouth detection results. As shown in figure 6, other parts can be detected when we do face detection. Figure 7 shows an example of the rotated face. The steps for finding out the angle are as follows,

Finding angle:

- Find the middle point of two eyes' boxes.
- Find the middle point of two points in step 1.
- Find the middle point of the mouse box.
- Connect the points found in step 2 and 3.
- Calculate the angle between the line in step 4 and vertical line.

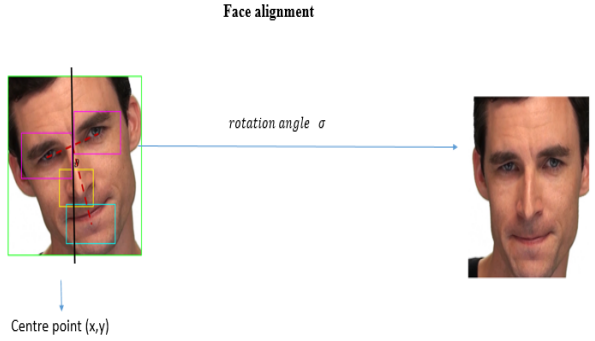

Figure 7. Example of a rotated face

If the angle of the rotated face can be found, then the straight face can be re-cut from the original image (before face detection).

3.2 Convolution Neural Network design

Convolution neural network (CNN) is a very powerful classification model. CNN extract features using convolution layers. Normal CNN process is shown in figure.8. In CNN structures there are some layers used in processing the data.

CNN layers:

- Convolution layer: A filter goes all over the image, multiply its elements with original image matrix, sum up to extract features from the original image. The output is called a feature map.
- Pooling layer: Uses the maximum value in the local area to represent this area, keeps the most important information and remove the rest, this can reduce the size of feature map and thus reduce computation costs.
- Fully connection layer: Connects all features and sends the output value to the classifier
- Softmax layer: Maps the output of multiple neurons to the interval of (0,1) which can be considered as a probability.

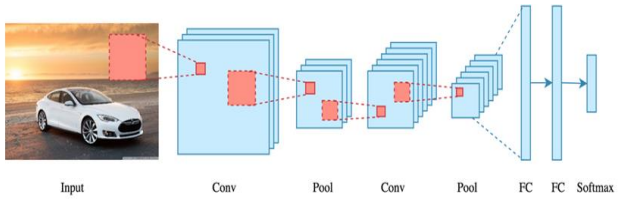

Figure 8. CNN classification process

Some CNN structures have shown pretty well classify ability, such as Alex-net, google-net, res-net. These structures have been built and tested in this system.

4. RESULTS AND ANALYSIS

The whole system is combined with a new pre-processing method and a well-trained CNN. The CNN structure is chosen from Alex-net, google-net and different deep Res-net structure. The best results of these three structures are used. The video database used in this paper is a video song in RAVDESS. This database includes six emotions which are happy, angry, neutral, calm, fear, sad. The original video frames are 1280*720, which is too big to train. So, we resized every frame after pre-processing to the size that can fit the structure input size. For instance, Alex-net structure input size is 227*227*3. Three is the RGB channels number. This dataset has 24 volunteers with 12 females and 12 males, each of them singing the same sentence with emotion in one video. Singing video set has 6 classes which are neutral, calm, happy, angry, fear and sad. The subtract gap is set as 5 and stride is set as 4 after several tests. After completion of all the pre-processing, the video set changes to a new image database and the detail of the database is shown in table 1.

Table 1. Details of database

Label	Count
Angry	5786
Calm	6644
Fear	5710
Happy	6021
Neutral	3024
Sad	6483

The database has 33668 images with six classes. This database is fed into CNN structure (Alex-net structure, google-net structure, and ResNet structure) and trained in Matlab. The training set, test set, and the validation set ratio is 8:1:1. ResNet can change the depth of the structure. ResNet-101, ResNet-51, ResNet-10, ResNet-8 and ResNet-4 are tested. After the test, ReNet-4 is chosen for comparison with other CNN structure, which can be well trained and have the best result of all ResNet structure.

Table 2. The different parameter values chosen for training the networks

Settings	Learning Rate	Max Epochs	Learning Rate factor	Laerning Rate Period	Max Batch Size	Environment
Alex-Net Base	0.001	15	0.7	2	128	Single GPU
Google-Net Base	0.0001	10	0.2	2	8	Single GPU
ResNet Base	0.0005	6	0.2	1	32	Single GPU

Table 2 shows the important parameters which play a vital role for training the neural networks. For AlexNet, based architecture the learning rate chosen is 0.001, the maximum epochs are 15, the learning rate factor is 0.7, learning rate period is 2 and maximum batch size is 128. For GoogleNet based architecture the learning rate chosen is 0.0001, the maximum epochs are 10, the learning rate factor is 0.2, learning rate period is 2 and maximum batch size is 8. For ResNet based architecture the learning rate chosen is

0.0005, the maximum epochs are 6, the learning rate factor is 0.2, learning rate period is 1 and maximum batch size is 32. Training of the neural networks done with the help of single GPU (NVIDIA GeForce GTX 1070).

Table 3. Accuracy Comparison

Authors	Data type	classifier	Accuracy
Biqiao Z et al	Acoustic+ Visual	Shared models	83.15%
Tuanbo G et al	Acoustic	Global feature SVM	79.40%
Frank A. Russo et al	Acoustic+Visual	247 raters	80%
Frank A. Russo et al	Visual	247 raters	75%
Frank A. Russo et al	Acoustic	247 raters	60%
Our model	**Visual**	**CNN**	**79.74%**

Table 3 shows the comparison between the accuracy rates achieved using different datasets and classifiers. There are two classifiers that achieved better accuracy rate than our model which are Biqiao Z et al. model based on shared models classifier which achieved an accuracy rate of 83.15% on Acoustic+Visual dataset and Frank A. Russo et al. model based on 247 raters classifier achieved an accuracy rate of 80% on Acoustic+Visual dataset. However, on testing only visual data type our model is able to produce the best accuracy rate of 79.74% using AlexNet structure.

In figure 9 the blue line shows the training accuracy and red line shows the loss of the cross-entropy in training set. The black line linked by some points is validation accuracy and loss. In the AlexNet training process, we found the validation loss is smooth after 4 to 5 epochs iteration. But the training accuracy remains higher, which means the model starts overfitting. So we stop here and get a 79.74% accuracy.

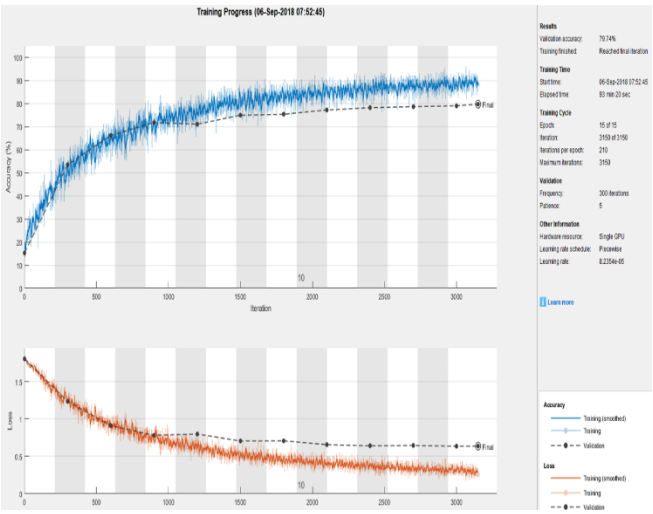

Figure 9. AlexNet structure training process

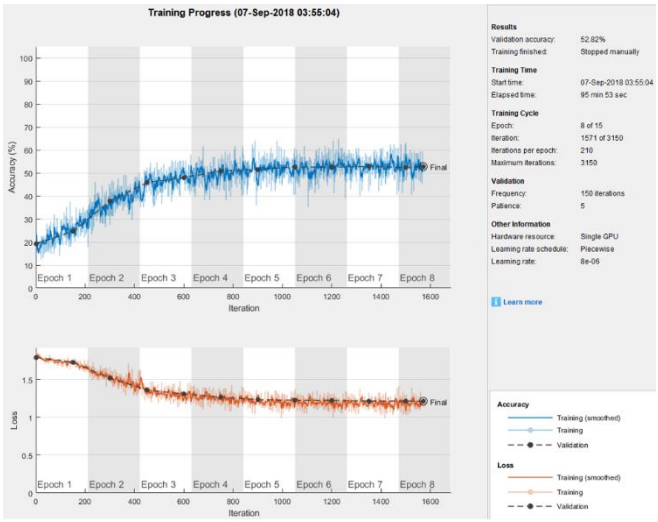

Figure 10. GoogleNet structure training process

Figure 11. ResNet structure training process

The Resnet structure is well trained as shown in the figure 11. The loss is very small and learning rate is also very small which means this model fits the dataset very well, but the accuracy is 75.89% at the end. The learning ability of this model is not good enough, a deeper structure might lead to a higher accuracy.

Table 4. Results of three structures

CNN Structure	Accuracy(%)
Alex net structure	**79.74**
Google net structure	62.89
ResNet-4	75.89

From table 4 it is clear that the overall AlexNet based structure and ResNet based structure show the best result, but GoogleNet has over 100 convolution layers, while AlexNet and ResNet-4 have only 12 convolution layers. Deeper ResNet structure has also been tested, but the performances are poor. Normally, deep convolution structure can extract features and can understand picture better. However, the pre-processing operation reduces many features which are not that important and cause the most static pixel to become black. The new image dataset created in this method is not complex to learn because most of the meaningless points change to

zero. In some ways, pre-processing operation helps CNN understanding the pictures. Figure 12(a) and 12(b) show some real test of this model.

Figure 12. (a) Real test (camera vision: angry, calm, sad, happy, fear and neutral)

Figure 12. (b) Real test (difference vision: neutral, angry, happy sad, calm)

5. CONCLUSIONS

The proposed pre-processing method can produce 79.78% accuracy which is 4.78% higher than the accuracy rate registered by humans. The method can analyze people's emotions, but the frame per second is not that big enough for smooth live video analysis. Using better face detection and face alignment technique can lead to a more accurate result. However, there is a need for verification of the algorithm or processing method. One of the drawbacks of this method is that only straight way face can be detected and analyzed because this method needs straight face features to analyze the emotion. If the face is not well detected or covered then the results will be poor. Also, the dataset does not include side face videos.

6. ACKNOWLEDGMENT

Our deepest gratitude and thanks to University of Strathclyde, Glasgow for providing us the best hardware configuration and the licensed version of MATLAB R2018a for the accomplishment of our work presented in this paper.

7. REFERENCES

[1] B. H. Mohammad Mahoor. 2017. Facial Expression Recognition Using Enhanced Deep 3D Convolutional Neural Networks. pp. 30-40, 2017.

[2] J. Donahue *et al.*. 2015. Long-term recurrent convolutional networks for visual recognition and description. pp. 2625-2634, 2015.

[3] W.-S. Chu, F. De la Torre, and J. F. Cohn. 2017. Learning spatial and temporal cues for multi-label facial action unit detection. *12th IEEE International Conference on Automatic Face & Gesture Recognition,* pp. 25-32, 2017

[4] A. Graves, C. Mayer, M. Wimmer, J. Schmidhuber, and B. Radig. 2008. Facial expression recognition with recurrent neural networks. *Proceedings of the International Workshop on Cognition for Technical Systems,* 2008.

[5] B. C. Ko. 2018. A Brief Review of Facial Emotion Recognition Based on Visual Information. *sensors,* vol. 18, no. 2, p. 401, 2018.

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

[7] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler. 2015. Deep learning for emotion recognition on small datasets using transfer learning. *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pp. 443-449: ACM, 2015.

[8] D. H. Kim, W. Baddar, J. Jang, and Y. M. Ro. 2017. Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. *IEEE Transactions on Affective Computing,* 2017.

[9] A. Graves and J. Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks,* vol. 18, no. 5-6, pp. 602-610, 2005.

[10] M. Cannici, M. Ciccone, A. Romanoni, and M. Matteucci.2018. Event-based Convolutional Networks for Object Detection in Neuromorphic Cameras. *arXiv preprint arXiv:1805.07931,* 2018.

[11] S. R. Livingstone and F. A. Russo. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one,* vol. 13, no. 5, p. e0196391, 2018.

[12] J. Jeon et al. 2016. A Real-time Facial Expression Recognizer using Deep Neural Network. *Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication*, p. 94: ACM, 2016

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

[14] C. Szegedy et al. 2015. Going deeper with convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1-9, 2015.

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

[16] H. A. Rowley, S. Baluja, and T. Kanade. 1998. Neural network-based face detection. *IEEE Transactions on pattern analysis and machine intelligence,* vol. 20, no. 1, pp. 23-38, 1998.