A STUDY INTO HUMAN EMOTION DETECTION USING COMPUTER VISION AND MACHINE LEARNING

OCTOBER 2, 2019

STUDENT NUMBER: S1602293
University of Gloucestershire

Contents

INTRODUCTION

Justification

Emotional detection in computing is an essential field of research due to its many use cases such as "a device that assisted children with Asperger syndrome read and respond to facial expressions" stated by Schwartz (2019). Emotional detection research will lead to an improved state of interaction between humans and their devices and is already showing promise with how "emotion detection technology is now being used to monitor and detect driver impairment, test user experience for video games and to help medical professionals assess the wellbeing of patients" also stated by Schwartz (2019). Schwartz infers that emotion detection can be used heavily in the field of market research and that emotion detection is currently a 20 billion dollar industry. This market is also seen to be growing at a rapid rate as indicated by a key figure in the research field of emotional detection, Kaliouby in their 2015 TED talk.

Emotion detection in computing is a growing area of technology, and there is continuing research into this area. There is no established best algorithm, and this is backed up by Kerkeni *et al.* (2018) when discussing the research area being young. Some emotion detection algorithms have been seen to have some racial bias. Specifically, Rhue (2019) had found that certain implementations of emotion detection tended to assign emotions of anger, surprise and fear more often to black men's faces than that of white men's faces.

Aims

This paper aims to investigate the methods used in the field of human emotion detection, uncover the underling thoughts on why specific methods are used over others and try to provide a base understanding of how anyone can accomplish the production of a human emotion detector software.

Objectives

- Thoroughly review literature related to emotion detection to identify the main discrepancies in views on the best methods to use when making an emotion detection software.
- Explain the essential methods of emotion detection.
- Create an outline for the appropriate output example that demonstrates the best methods of emotion detection discussed in the paper.

LITERATURE REVIEW

Overview

The literature review will indicate how computers can detect the emotions of their users with different methods. These methods will be explained, compared and contrasted against other methods. The methods shown will be justified and alternatives to similar problems given.

What emotional indicators can be seen by computer vision?

Without the use of specialist hardware, the emotional indicator of scent cannot read in a significant sample size for machine learning; thus, the emotional indicator of scent is out the window. Boukis *et al.* (2007) discuss that due to cameras and microphones being commonplace on most devices, most computers are capable of detecting facial gestures, speech, body gestures and movement. The easiest of these to use for emotion detection purposes will be facial expressions and speech due to the abundance of data available for these indicators and the inherent lack of data on body gestures and movement and their relation to emotions. The benefit to having these hardware features of microphones and cameras on a plentiful amount of systems is that it provides a large sample size as to ascertain an accurate, measurable data set.

How can the emotional indicator of speech be used as data to produce an emotional verdict by way of computer vision and machine learning?

Summary

To be able to detect human emotion from speech, three things are needed. Emotional speech data, a way to extract features of audio, and a way to classify the features by way of a machine learning algorithm. These three factors work in unison. The data gets its features extracted. An example of feature extraction is retrieving the pitch of the speech. The features then get put into a machine learning algorithm and told by the data what emotion those features represent, thus training the algorithm. Data not used to train the algorithm can then be passed through it to test its accuracy by withholding the label of the data indicating it is correct emotion, therefore, making the algorithm reach an informed decision on what emotion the data is conveying. After the training and testing of the algorithm, data collected using computer vision can be used to give the emotion of the user.

Collecting input

The input collection for emotional speech recognition is straightforward to gather as all that is needed is to record someone speaking and save it. Numerous devices are capable of recording speech, including phones, laptops, and tape recorders. The feature extraction should digitally process the saved speech signal and then have the features extracted passed into the classifier.

Emotional speech data sets

An emotional speech data set is used to train the algorithm. It must consist of a vast amount of emotional voice lines that are all labelled with what emotions are in the speech. Pitterman *et al.* (2010) have discussed how there are many of these emotional speech data sets already available to the public and in many different languages.

To pick a data set to use one should first decide on the audience of the output for the research for example if one is using the output on mainly English speakers then an English dataset will be most suitable, along with what emotions one wishes to record as if one wants data for happy, angry and sad speech then a data set without sad will be irrelevant for your purposes. Maheswari (2018) has stated that the most crucial factor is a data set with the highest quantity of data needed to make machine learning algorithms function their best.

The output for this paper focuses on English speakers; thus, the paper will continue to show to process of picking an English emotional speech data set. There are many English emotional speech data sets most notable are the Toronto Emotional Speech Set (TESS), the CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). The TESS database only who two different talkers and are both female this may lead to trouble as the algorithm may not be able to recognise the emotion of people with a different timbre of voice. The CREMA-D is the largest of the three databases, and one may wish to use it for that regard. However, the database this paper will use is RAVDESS its large amount of variety in its actors and the fact that it covers eight different emotions.

Feature extraction

Since emotion detection methods do not use the meaning of the words spoken in the detection and the fact that audio will contain features about who is speaking which, passing the raw audio into a classification will yield little to no results in emotion detection. Therefore features of the audio that indicates emotion must be extracted to use within the classification algorithm. Rajvanshi (2018) has talked about the two major types of audio feature extraction. The first type is local, and this accounts for all the features extracted each frame (intervals of 20-40 ms). The second type of feature is global, and these features are calculated based on the whole audio file.

Some global features extracted from audio speech files include modulation spectral features which form into a Modulation Spectrogram (MS), energy, pitch, articulation rate and qualitative features like harshness, tenseness and breathiness to name a few indicated by Sezgin, Gunsel and Kurt, (2012). They also show that some local features extracted include Mel-frequency Cepstral Coefficients (MFCCs) which make up a Mel-Frequency

Cepstrum(MFC), Log Frequency Power Coefficients (LFPC) and Linear predictive coding (LPC).

The features that are most indicative of emotion is a big topic of conversation in itself. Many researchers disagree about which features are the most suitable for speech emotion detection (Kerkeni et al., 2018). The uncertainty may be since certain features can only identify certain emotions and not all of the ones required to account for all the underlying emotions, as stated by Paul Ekman. In practice, this means to find the most suitable features to use; one would need to use every combination of features with every combination of dataset and classification. However, in the spirit of saving time and getting accurate enough results, this paper will continue with just one feature: Mel-Frequency Cepstral Coefficients (MFCCs). A way to get increased accuracy from the results would be to use a combination of features. Modulation spectral features which make an MS used in combination with MFCCs cover the identification of all emotions needed and are proven to be able to have an accuracy of above 90% when used together shown by Kerkeni et al., (2018).

The shape of the vocal track determines human speech. If one can state the shape of the vocal track, one can get a representation of the sound coming out. The short time power spectrum of the speech indicates the shape of the vocal tract. MFCCs represent this short-time power spectrum. To calculate MFCCs, one must first separate the speech signal into shorter length signals; this is because audio signals are continuously changing, and thus the length is shortened, to have a signal that does not change as frequently. A single frame of shortened audio is around 20-40ms. After separating the frames, the periodogram estimate of the power spectrum calculation for each frame takes place. The periodogram tells us what frequencies are within a frame. To compute the periodogram, use the following formula:

$$P_i(k) = \frac{1}{N}|S_i(k)|^2$$

Figure 1: Periodogram formula.

The Discrete Fourier Transform of the frame is needed to complete this calculation which is given by :

$$S_i(k) = \sum_{n=1}^{N} s_i(n)h(n)e^{-j2\pi kn/N} \qquad 1 \leq k \leq K$$

Figure 2: the Discrete Fourier transform formula.

In the formulas 'i' represents the frame number, 'Si(n)' would then be the framed signal, K is the DFT length, and h(n) is an N long analysis window. Since the audio file will contain every frequency, the microphone can pick up 140,000 different frequency's from 1Hz to 140kHz the frequencies will need to be grouped to allow for easier processing. To group frequencies together use a Fast Fourier transform(FFT) which will split the 140,000 frequencies into 512 groups. Out of the 512 groups, only the first 257 should be kept this is due to the range of human hearing as humans cannot hear frequencies higher than 28kHz and the 255 encompasses frequencies out of this range.

The next step is to apply a Mel-spaced filter bank to the periodogram power spectral estimate. After the application of the filter bank, 26 numbers indicating the energy of each filter bank remain. The filter bank energies should then be compressed using their logarithm as by doing this; the energies will better resemble what humans hear as humans cannot perceive loudness on a linear scale. The compression leaves us with 26 log filter bank energies that can be used to get 26 cepstral coefficients by computing the Discrete Cosine Transform. Coefficients that show fast changes can diminish performance and thus only the DCT coefficients 2-13 are kept leaving us with 12 out of 26 coefficients these are called the Mel Frequency Cepstral Coefficients. Delta coefficients allow more features extraction from audio, delta coefficients can be calculated by:

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2}$$

Figure 3: the formula for delta coefficients.

In total, there are 12 coefficients and 12 delta coefficients, which in turn produce a vector of features of length 24. (Practicalcryptography.com, 2019)

Engelbart (2018) says that although this type of feature extraction is suitable, it is also highly costly due to the computational power required to analyse the audio and run all the calculations. The major problem with audio analysis would be that to have real-time analysis; one would need to have as few pre-processing steps as possible (feature extraction). The use of the DeepEmoNet model mitigates this problem as it specialises in real-time emotional audio analysis.

Classification algorithm

A classification algorithm takes input data, in this case, the 24 feature vector and maps it to a category which in this case is a label of emotion by using the knowledge it has got from analysing training data as told by Edureka (2019). Edureka (2019) also reveals that there are many different classification algorithm types including but not limited to Linear Classifiers, Support vector machines, Kernal estimation, decision trees and Neural networks.

The big three classification algorithms that get used for auditory emotional detection are Multivariate Linear Regression (MLR), Support Vector Machine (SVM) and Recurrent Neural Networks(RNN). Kerkeni et al. (2018) show that the most accurate results from the three can be seen using a Recurrent Neural Network due to the suitability of using RNN's for learning time series data.

The RNN uses the same parameters for each step of the process, whereas a regular neural network uses different parameters. One downside to using RNN's is that they have vanishing gradients meaning that they cannot find long-range features in the data. A way to counter this would be to add Long Short Term Memory (LSTM) to the RNN, as explained by Thapliyal (2019). Chen and Jin (2019) explain that LSTM is the process of adding memory that persists throughout the whole analysis to enable long-range dependency exploitation.

The diagram below represents an RNN and its forward computation:

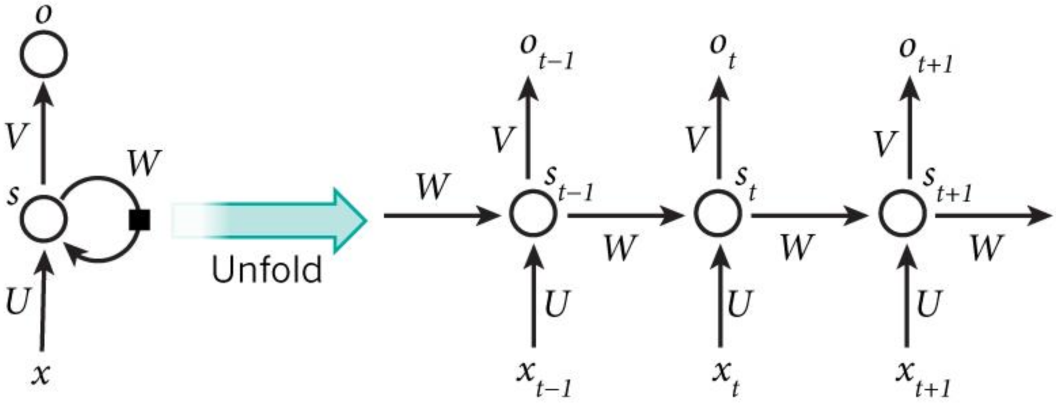

Figure 4: an RNN and its forward computation.

U, V and W are parameters matrices that stay the same throughout the network. X is input, S is the hidden state, t is the time step, and O is the output. The calculation for the hidden state formulas and variables goes as follows:

$$s_t = f(Ux_t + Ws_{t-1})$$

Figure 5: the formula for calculating the hidden state formulas and variables.

(Lim, Jang and Lee, 2016)

How can the emotional indicator of facial gestures be used as data to produce an emotional verdict by way of computer vision and machine learning?

Summary

There are three steps to get an emotional verdict from a facial gesture. The first is to use computer vision to get the input of the facial gesture. The second step is to train a classification model with an emotional face database. The third is to pass the input through the classifier to receive an output which, in this case, will be the emotion the facial gesture is representing. This segment of the paper will be explaining all these steps but with an emphasis on how to get a verdict in real-time due to not being able to get it from the audio analysis method.

Collecting facial gesture input with computer vision

What inputs are needed to classify an emotion from a facial gesture? The inputs needed is a face but not just a face. The face has to be 48*48 pixels in size and to be in greyscale as that is what the classifier is going to be trained to take as input but more on that later. Acquiring a face can be done through the use of computer vision and object detection. Kumar (2019) indicated that one of the fastest methods compared to other face detection methods is Haar feature-based cascade classifiers.

Viola and Jones (2019) explain that a Haar cascade takes a bunch of positive images (images with faces) and negative images (images without faces) and finds out what features make up a face by using rectangles and summing up the pixel intensities(of the greyscale image) within it to find the difference between these sums. These Haar features are remembered in stages meaning that if a region of an image has the first Haar features it then tries to apply the next set of features to see if the image still qualifies as a face. A face region then moves around the image, trying to apply the stages of features until it finds all faces.

OpenCV has an implementation of Haar cascades that this paper will use. The OpenCV library can also handle all of the pre-processing of the image. It has functions to capture an image from the built-in camera, to convert the image to greyscale for the Haar Cascade classifier and can resize the bounded face to 48*48 ready for the emotion classification shows Bradski (2000). OpenCV even provides a dataset for the Haar Cascade classifier specifically for facial detection, which is a huge timesaver. An example of the OpenCV library Haar Cascade classifiers in action is in the figure below:

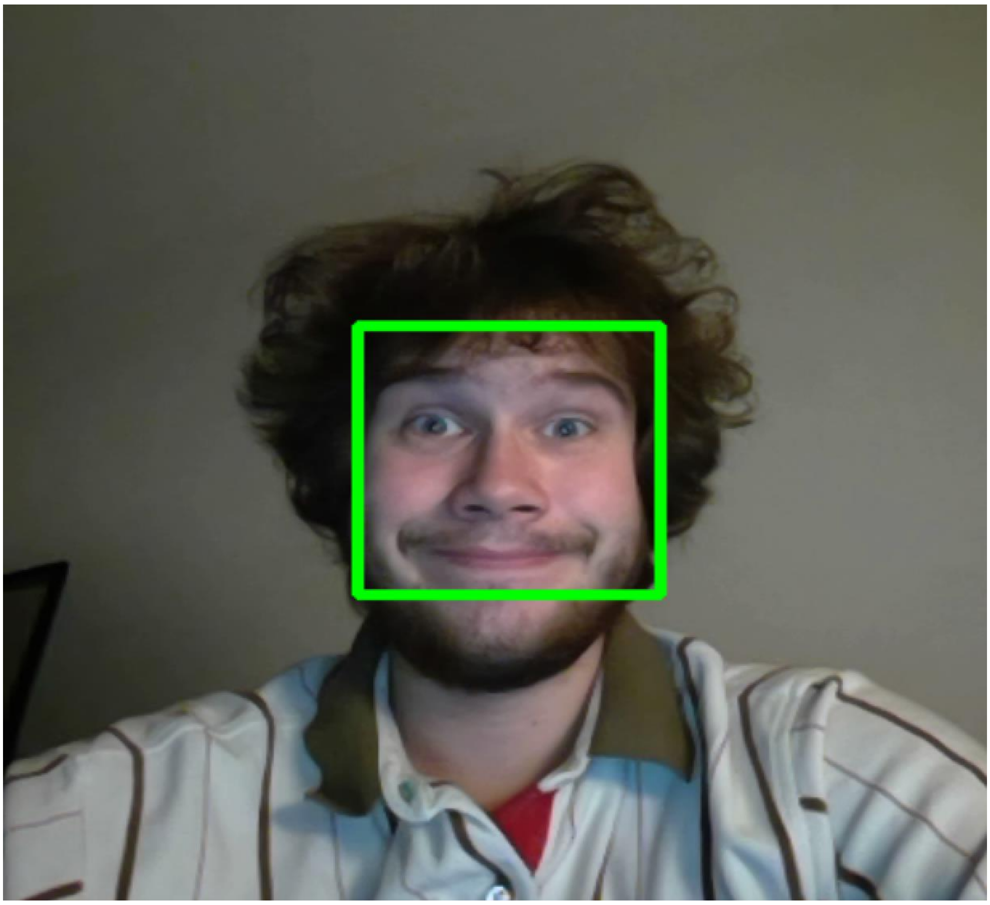

Figure 6: a bounding box around an identified face.

Emotional Face Datasets

There are many datasets for emotional faces to choose. One could even make a dataset themselves, but that would take a long time, in light of that fact, this paper will go through the process of picking the most relevant dataset. The website Face-rec.org (2019) instructs that when creating a new algorithm for classification, one should pick a standardised dataset to compare results with other researchers such as the Color FERET Database, USA. Since creating a new algorithm for classification is not the goal of this paper the focus of the search should be on how accessible the database is, how many emotions are in the database and how extensive the database is.

One free dataset that satisfies all the search constraints would be the dataset created for a Kaggle.com (2019) challenge aptly named"
Challenges in Representation Learning: Facial Expression Recognition Challenge". The paper will continue to reference this dataset as the FER dataset. The FER dataset contains 35,888

samples which satisfy the quantity constraint. The dataset also encompasses seven different emotions Angry, Disgust, Fear, Happy, Sad, Surprise and Neutral. The data is a CSV file that consists of two columns: one being a sequence of floats indicating each pixels greyscale intensity for all 48*48 pixels of the image and the other being an integer indicating the category of which the image belongs for example 3 which indicates the image shows a happy face. Considering that the dataset is ample, free, encompasses many emotions and is simple to use due to the format of the CSV file, this paper will use this dataset for the output design.

Classification by way of a CNN

The Algorithmia Blog (2019) shows that the use of Neural networks is accepted to be the best way to detect emotion from imagery with computers specifically Convolutional Neural Networks (CNNs) as they are made to make use of images as inputs. Saha (2019) says that CNN is a type of deep learning algorithm that gets an image as input and then assigns weights to objects/features it can detect within the image. CNN's can also understand how the object/features are different from each other, according to Saha (2019). Saha (2019) also points out that the fact that a CNN can detect relevant features by itself diminishes the need for handcrafted filters and minimises the pre-processing of the input substantially, therefore, reducing computation time.

How do CNNs work? Well, CNN's work similarly to how the visual cortex of the human brain works. Neurons get information from a small area of an image and these neurons overlap covering the whole image. The input image, in this case, is a 48*48 matrix of floats indicating the greyscale intensity of the associated pixel. CNN's work in two main steps: the first step is feature learning, and the second is classification. Both steps usually involve a multitude of layers. Layers found in the feature learning step include convolution and pooling but may also include rectified linear unit (ReLU). Layers in the classification step can include flatten, fully connected, and SoftMax, according to Cs231n.github.io (2019).

The convolution layer produces a feature map that shows the locations of features and how prominent they are by applying filters to the input. The filter is also known as a Kernal is a matrix that is smaller than the input matrix that goes along a segment of the input matrix from left to right then moves down and goes from left to right again until the whole segment has been filtered to produce convolved features as Brownlee (2019) shows. The figure below demonstrates this process.

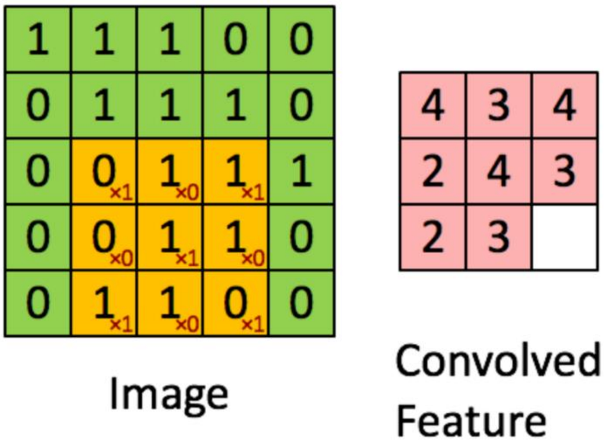

Figure 7: Convoluting a 5x5x1 image with a 3x3x1 kernel to get a 3x3x1 convolved feature

A pooling layer is similar to a convolution layer, but the main focus of the pooling layer is to reduce the size of the convolved feature. Pooling is beneficial as less power is needed to compute to process the data. The pooling is also useful for extracting dominant features that are never rotated and never change position. The figure below demonstrates two types of pooling Average and Max pooling:

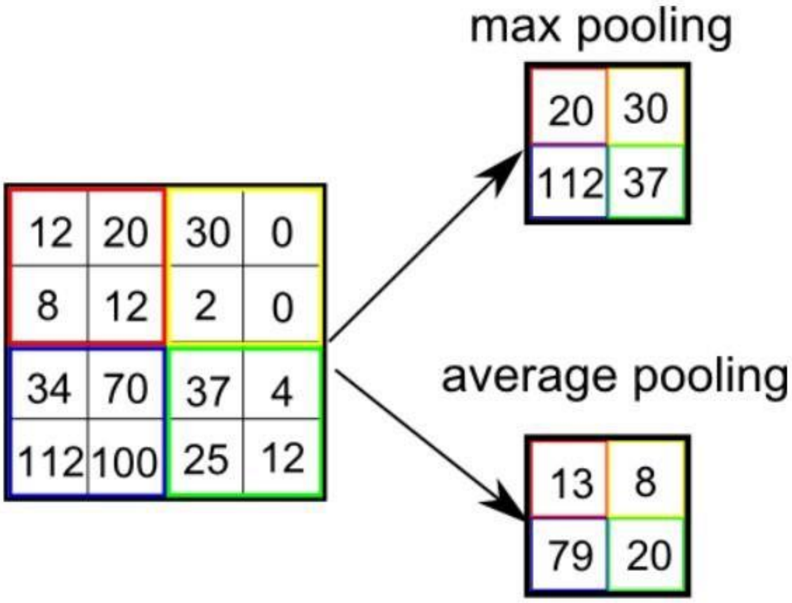

Figure 8: Types of pooling

Max-pooling takes the max values, and average pooling takes an average. Max pooling is the preferred method for pooling as it outperforms average pooling by also acting as a Noise Suppressant says Saha (2019). Google Developers (2019) state that Softmax is a layer in the classification step of CNN's. Softmax gives an output of probabilities that the image could be representing, for example, 5% neutral and 95% happy.

CNN models can vary significantly from what layers are used to how many and in what order. Arriaga, Valdenegro-Toro and Plöger proposed one modern model for emotion classification in 2019 called Mini Xception, shown in the figure below. The model moves away from most CNN models as it does not use a fully connected layer at the end. The model is also different as it uses residual modules and depth-wise separable convolutions which have been seen to be very successful at allowing for state-of-art performance as shown by Kumar (2019). Therefore this papers output will be using this model although due to using a SoftMax layer at the end of the CNN sometimes there will not be a definitive emotion that the image of a face is conveying if the probabilities of being a particular emotion are high in more than one emotion and thus a proposed solution to this is going to be in the output design of the paper.

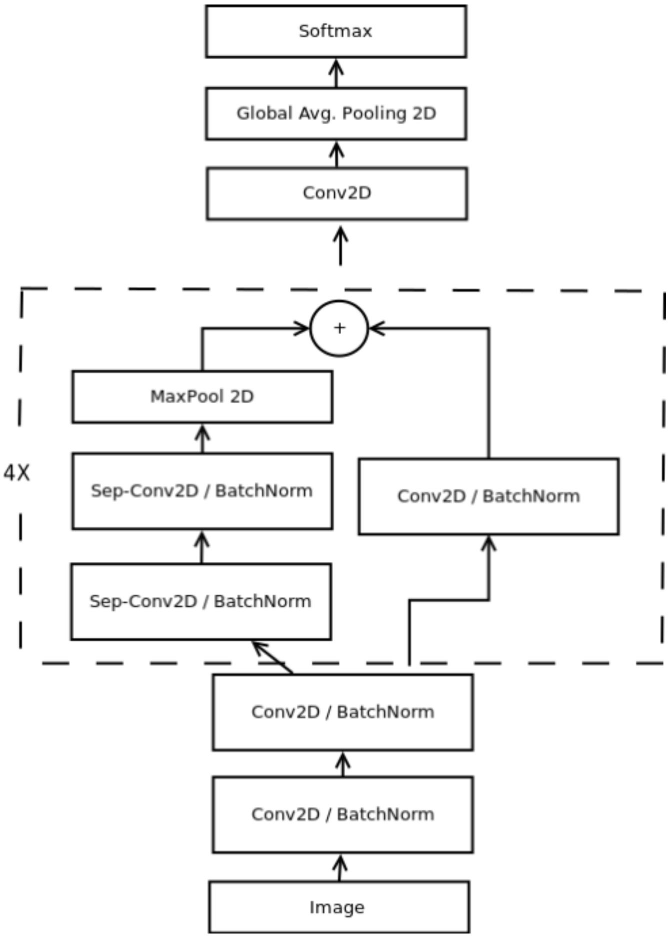

Figure 9: Proposed mini Xception model for emotion classification.

Summary

In summary, an automated emotion detection method is a powerful tool, but choosing the correct steps to achieve such a feat could make or break the output. Hopefully, this paper has been informative into what the correct steps are for different outputs. Emotion detection still has a long way to come to be able to deliver higher accuracy and certainty results.

OUTPUT DESIGN

The find output of this paper will be a videogame where the sole input is what emotion the user is perceived to be showing. Considering how inputs for games are usually relatively lag-free (calculated in real-time) the only emotion detection the game will use will be from facial input as the audio analysis methods proposed in the paper are meant for pre-recorded audio and cannot calculate a verdict in real-time.

Since the game will now have the sole input of faces the output from the CNN will have to yield a definitive result, and as since it currently produces an array of probabilities for each emotion some calibration may be needed. An example of such a calibration could be subtracting the neutral emotion results from the user from the other emotions only to see the changes in emotion. Another calibration method could be to take the output over a time frame and add the results up to give the most prominent emotion over a time frame. A final solution would be to add multipliers to the percentages this could help if the CNN has a hard time assigning higher percentages to certain emotions like disgust for example when making a disgusted face the angry percentage and disgusted percentage both increase but if a multiplier were to apply to the disgusted percentage then it would appear to be the most prominent emotion. The multiplier may even just be applied during the actual game to give some leeway as the output is only 66% accurate.

The first step in making the output would be to gather emotional facial input. The OpenCV library handles the input and preprocessing. The second will be to use a pre-trained CNN model trained to recognise emotions using the model proposed. Figure 10 is a visualisation of the output of such a model.

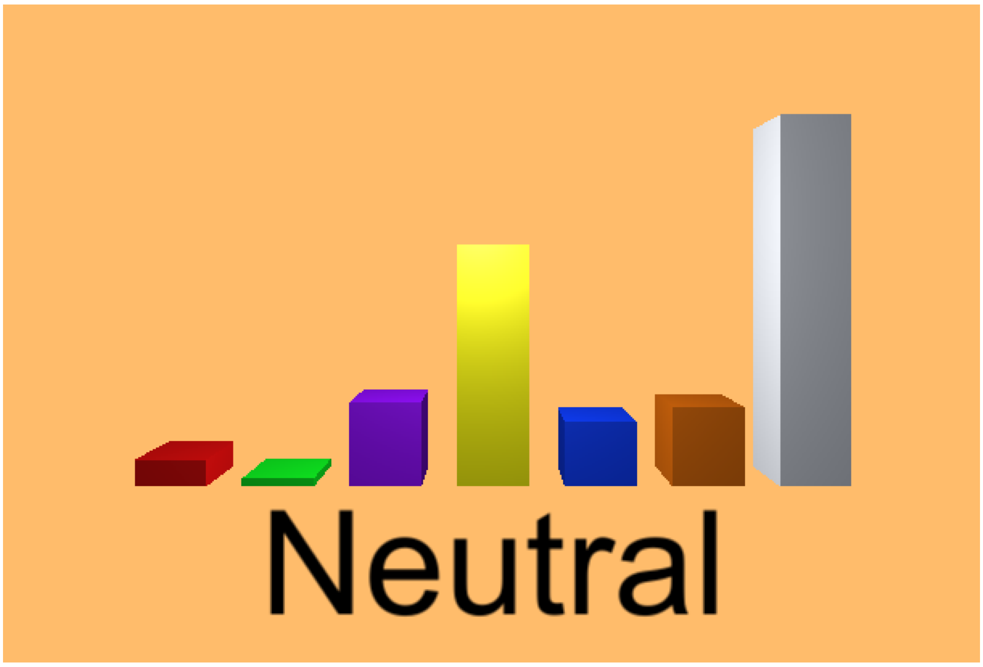

Figure 10: the output of the CNN visualised with the most prominent emotion indicated.

The game will be created using Unity, a game creation engine that uses C# scripts. OpenCV is a python only library, and therefore a workaround must be applied. Trying to get an

OpenCV wrapper for C# working with unity itself can get a bit hairy not to mention adding Keras/Tensorflow wrappers which are the libraries also made in python for deep learning and neural networks. The solution this paper will use is to make the CNN using python then use Pyinstaller(a python library that allows the bundling of python code and all its dependencies into an application) to create an application to be executed within Unity. The application should output the emotional verdicts to a file that unity can then read.

The third step of this game creation is to make a fun game that is designed around using emotions as input. The main goal of the game would be to react accordingly to different stimuli at an ever-increasing pace. The setting where this game is going to take place is a kind of a dating simulator based at restaurant table at a speed dating convention, the user would need to react to how their date acts for example if their date is crying then the user must act sad and if their date angry then the user must also act angry. After going through all the dates in the round, the user receives an indication as to how many people said they would go on a second date with them in accordance to how they reacted at the time. The rounds will go faster and faster meaning the user has less time to react per date until the user has failed by getting no dates and the number of rounds partaken in will be the user's score. The dates will be the one coming too and from the table as too minimise set design. The game will be a 2D based game with hand-drawn faces and animations.

REFERENCES

- Algorithmia Blog. (2019). *Introduction to Facial Emotion Recognition | Algorithmia Blog*. [online] Available at: https://algorithmia.com/blog/introduction-to-emotion-recognition [Accessed 3 Nov. 2019].
- Arriaga, O., Valdenegro-Toro, M. and Plöger, P. (2019). *Real-time Convolutional Neural Networks for Emotion and Gender Classification*. [online] arXiv.org. Available at: https://arxiv.org/abs/1710.07557 [Accessed 4 Nov. 2019].
- Boukis, C., Pnevmatikakis, A. and Polymenakos, L. (2007). *Artificial Intelligence and Innovations 2007: from Theory to Applications*. Boston, MA: International Federation for Information Processing, p.247.
- Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

- Brownlee, J. (2019). *How Do Convolutional Layers Work in Deep Learning Neural Networks?*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/convolutional-layers-for-deep-learning-neural-networks/ [Accessed 3 Nov. 2019].
- Cassidy, R. and Smith III, J. (2019). *What is a Filter Bank?*. [online] Ccrma.stanford.edu. Available at: https://ccrma.stanford.edu/realsimple/aud_fb/What_Filter_Bank.html [Accessed 15 Oct. 2019].
- Cs231n.github.io. (2019). *CS231n Convolutional Neural Networks for Visual Recognition*. [online] Available at: https://cs231n.github.io/convolutional-networks/ [Accessed 3 Nov. 2019].
- Edureka. (2019). *Classification Algorithms | Types of Classification Algorithms | Edureka*. [online] Available at: https://www.edureka.co/blog/classification-algorithms/ [Accessed 1 Nov. 2019].
- Engelbart, J. (2018). *A Real-Time Convolutional Approach To Speech Emotion Recognition*. Master. The University of Amsterdam.
- Face-rec.org. (2019). *Face Recognition Homepage - Databases*. [online] Available at: http://www.face-rec.org/databases/ [Accessed 3 Nov. 2019].
- Google Developers. (2019). *Multi-Class Neural Networks: Softmax | Machine Learning Crash Course*. [online] Available at: https://developers.google.com/machine-learning/crash-course/multi-class-neural-networks/softmax [Accessed 3 Nov. 2019].
- Kaggle.com. (2019). *Challenges in Representation Learning: Facial Expression Recognition Challenge | Kaggle*. [online] Available at: https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/overview [Accessed 3 Nov. 2019].
- Kaliouby, R. (2015). In: *This app knows how you feel — from the look on your face*. [online] Available at: https://www.ted.com/talks/rana_el_kaliouby_this_app_knows_how_you_feel_from_the_look_on_your_face/up-next?language=en [Accessed 8 Oct. 2019].

- Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K. and Mahjoub, M. (2018). Speech Emotion Recognition: Methods and Cases Study. *Proceedings of the 10th International Conference on Agents and Artificial Intelligence*. [online] Available at: https://www.researchgate.net/publication/322873355_Speech_Emotion_Recognition_Methods_and_Cases_Study/stats [Accessed 9 Oct. 2019].
- Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K. and Mahjoub, M. (2018). Speech Emotion Recognition: Methods and Cases Study. *Proceedings of the 10th International Conference on Agents and Artificial Intelligence*.
- Kumar, A. (2019). *Demonstration of Facial Emotion Recognition on Real Time Video Using CNN : Python & Keras - Machine Learning in Action*. [online] Machine Learning in Action. Available at: https://appliedmachinelearning.blog/2018/11/28/demonstration-of-facial-emotion-recognition-on-real-time-video-using-cnn-python-keras/ [Accessed 3 Nov. 2019].
- Kumar, A. (2019). *Demonstration of Facial Emotion Recognition on Real Time Video Using CNN: Python & Keras - Machine Learning in Action*. [online] Machine Learning in Action. Available at: https://appliedmachinelearning.blog/2018/11/28/demonstration-of-facial-emotion-recognition-on-real-time-video-using-cnn-python-keras/ [Accessed 4 Nov. 2019].
- Lim, W., Jang, D. and Lee, T. (2016). Speech Emotion Recognition using Convolutional and Recurrent Neural Networks. In: *APSIPA ASC 2016*. [online] Daejeon: Audio and Acoustics Research Section, ETRI, p.2. Available at: http://www.apsipa.org/proceedings_2016/HTML/paper2016/137.pdf [Accessed 2 Nov. 2019].
- Maheswari, J. (2018). *Breaking the curse of small datasets in Machine Learning: Part 1*. [online] Medium. Available at: https://towardsdatascience.com/breaking-the-curse-of-small-datasets-in-machine-learning-part-1-36f28b0c044d [Accessed 8 Oct. 2019].
- Pittermann, J., Pittermann, A. and Minker, W. (2010). *Handling emotions in human-computer dialogues*. Dordrecht: Springer Science+Business Media B.V.

- Practicalcryptography.com. (2019). *Practical Cryptography*. [online] Available at: http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/ [Accessed 1 Nov. 2019].
- Rajvanshi, K. (2018). An Efficient Approach for Emotion Detection from Speech Using Neural Networks. *International Journal for Research in Applied Science and Engineering Technology*, 6(5), pp.1062-1065.
- Rhue, L. (2019). *Emotion-reading tech fails the racial bias test*. [online] The Conversation. Available at: https://theconversation.com/emotion-reading-tech-fails-the-racial-bias-test-108404 [Accessed 9 Oct. 2019].
- Saha, S. (2019). *A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way*. [online] Medium. Available at: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53 [Accessed 3 Nov. 2019].
- Schwartz, O. (2019). *Don't look now: why you should be worried about machines reading your emotions*. [online] the Guardian. Available at: https://www.theguardian.com/technology/2019/mar/06/facial-recognition-software-emotional-science [Accessed 8 Oct. 2019].
- Sezgin, M., Gunsel, B. and Kurt, G. (2012). Perceptual audio features for emotion detection. *EURASIP Journal on Audio, Speech, and Music Processing*, 2012(1).
- Szeliski, R. (2011). *Computer Vision*. London: Springer-Verlag London Ltd.
- Thapliyal, M. (2019). *Vanishing Gradients in Recurrent Neural Networks*. [online] Medium. Available at: https://medium.com/mlrecipies/vanishing-gradients-in-recurrent-neural-networks-b231c2afde4 [Accessed 2 Nov. 2019].
- Viola, p. and Jones, M. (2019). Rapid Object Detection using a Boosted Cascade of Simple Features. In: *CVPR*. [online] Available at: https://www.cs.cmu.edu/~efros/courses/LBMV07/Papers/viola-cvpr-01.pdf [Accessed 3 Nov. 2019].

CITATIONS

- Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset. *IEEE transactions on affective computing*, *5*(4), 377–390. DOI:10.1109/TAFFC.2014.2336244
- Chen, S. and Jin, Q. (2019). *Multi-modal Dimensional Emotion Recognition using Recurrent Neural Networks*. [online] Diuf.unifr.ch. Available at: https://diuf.unifr.ch/main/diva/recola/data/AVEC_2015_Chen.pdf [Accessed 2 Nov. 2019].
- Dupuis, K., & Pichora-Fuller, M. K. (2010). *Toronto emotional speech set (TESS)*. Toronto, University of Toronto, Psychology Department.
- Flanagan, Patricia (January 25, 2011). "Face Recognition Technology (FERET)".
- Livingstone SR, Russo FA (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 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391
- Unity Technologies, 2019. *Unity*, Available at https://hadoop.apache.org.
- Chollet, F. (2015). Keras. *GitHub repository*. [online] Available at: https://github.com/fchollet/keras [Accessed 4 Nov. 2019].

GLOSSARY OF TERMS

- Computer vision **–** Human beings can perceive the complexities of the world with relative ease. While looking at any scene, one should be able to name the objects one sees, Identify the smells one senses and the sounds one hears. However, computers cannot inherently "interpret an image at the same level" (Szeliski, 2011) as humans. Therefore computer vision is a field of study in how to get computers to a high-level understanding of the world around them.
- Emotional indicator - An emotional indicator is anything that a human does that allows someone to infer their emotion. For example, if someone smiles, then one could infer that that person is happy. There are audible, sensory and visual indicators along with unseen indicators like habit changes or personality shifts.