Facial Expression Recognition for Cloud Robotics

Charlie Maclean
University of Cambridge
Cambridge, United Kingdom
cm927@cam.ac.uk

ABSTRACT

As social robots become widespread, it will be essential that they can classify the emotions of humans around them, in order to interact in a meaningful and helpful way. But limited hardware means they may have to offload video data to the cloud, reducing the resolution of the content. This work focusses on evaluating the fitness of neural network models for cloud computing, specifically the effect of changing spatial and temporal resolution on the classification of emotion in video. I build different models to asses each models performance when given lower resolution video. The results show that by applying a CNN-LSTM model to a 8-class problem, we can achieve 81% accuracy on high resolution video, and maintain 76% accuracy at the lowest resolution. To the best of my knowledge this is the first work that investigates the effect of changing both spatial and temporal resolution on video-based sentiment classification.

KEYWORDS

Affective computing, robotics, cloud computing, emotions, arousal, valence, resolution.

ACM Reference Format:
Charlie Maclean. 2021. Facial Expression Recognition for Cloud Robotics. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Social robots are becoming increasingly widespread, with uses in a wide range of locations, providing help in hospitals [5], care homes [7] and schools [26]. These robots are frequently required to interact meaningfully with humans, and in order to do so it is essential that they are able to classify emotions to react accordingly. However, many social robotic platforms lack the computational hardware required to perform classification [3]. Hence, it will likely become necessary to move to a cloud robotic framework, where sensing data is offloaded to the cloud and processed there. Unreliable network conditions mean we must be prepared for video data to enter the cloud at reduced spatial and temporal resolutions.

In past years, neural networks have become ubiquitous for classifying emotion, due to their ability to learn pattern humans would be unable to program in. However they can suffer from not being generalizable, especially if a network is trained in one domain, then deployed in another. For example, a network trained on high resolution data would be less effective at classifying low resolution data.

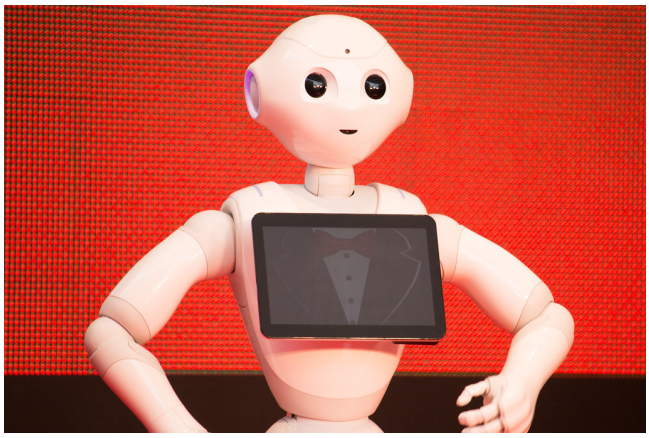


Figure 1: Pepper by SoftBank Robotics, an example of a social robotic platform which has been trialled in care homes in the UK. Photograph by Dick Thomas, via creative commons (source).

For classifying images, there is a large volume of work looking at using convolutional neural networks (CNNs). When applied to an image, a CNN convolves a filter with the pixel data, generating meaningful features. CNNs have found widespread use across various domains, including facial recognition [21] and object detection and classification [34]. Several architectures have been proposed offering impressive ability to learn features from video using purposes, for example the VGG16 network [32] and the ResNet50 network [13], both of which were able to achieve winning results in the ImageNet object detection and classification challenge [30].

For classifying videos, it is often vital to take temporal data into account, and as a result Recurrent Neural Networks (RNNs) [29] are a good choice. RNNs have some internal state, or memory, which they use to process sequences, learning patterns that may vary over time. A very popular architecture is Long Short Term Memory [14] (LSTM), which make use of gates to control the flow into and out of cells in the architecture. LSTMs have found wide usage, across speech recognition [11], market prediction [31] and handwriting recognition [10].

In this work we create a couple of classifiers which are tested on video at a variety of resolutions and frame-rates, in order to deduce which classifiers may be useful for cloud robotics. The classifiers are tested on a 7-class video dataset, and our results show that we can achieve an accuracy of X on

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Woodstock '18, June 03–05, 2018, Woodstock, NY
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

The rest of the paper is structured as follows: Section 2 gives an overview of previous work on similar problems, Section 3 gives detail about the methodology employed in the study. Then, Section 4 discusses the results before Section 5 goes over future research directions. Finally, Section 6 concludes the paper with an overview of the findings.

2 RELATED WORK

In this section I give an overview of facial expression recognition techniques, first for images, and then for videos, followed by a section on classifying data at reduced resolutions.

2.1 Facial Expression Recognition in Images

Facial expressions have allowed humans to communicate their emotions amongst each other for years. There has been a large volume of research into the mechanisms which allow this non-verbal communication to happen. An early work by P. Ekman and W.V. Friesen introduced the Facial Action Coding System (FACS) [8], which described a list of facial action units - regions which change as a person changes their expression. The work further describes how a given facial expression could be described as a combination of action units. Following this work, P. Ekman et al. detailed how a mapping can be made between the facial action units and a person's emotions [6]. I will now detail the work that has been put into using machines to detect emotions from photos, splitting the research into those that make use of deep networks, and those before deep networks became widespread.

Before the use of deep networks, there were two main approaches to sentiment classification for images - rule-based methods, and appearance based methods. First, the rule-based methods were centred around detecting facial action units individually, and then piecing together the results from the facial action unit recognizers to derive an overall emotion, for example in Y. Tian et al.'s work [35]. These techniques suffered as recognizing an individual action unit is not an easy task. In the alternative approach, appearance based methods, some features are extracted from the overall face, and then those features are passed through a machine learning classifier. For example, M.S. Bartlett et al. found that they could achieve good results by extracting features using Gabor filters, followed by a Support Vector Machine (SVM) [2]. Additionally, J. Whitehall and C. Omlin showed that comparable results could be obtained in significantly quicker time by instead using Haar wavelets to obtain features before passing through a SVM [40].

In previous techniques, processing of emotions had been split into learning features, selecting features and then using a classifier to learn the patterns. The downside of taking that approach is that the first layers do not get feedback from the latter layers. Deep learning aims to solve that, by integrating the feature finding, selection and classification into one deep network that can be trained at once. One of the early papers making use of this technique was by P. Liu et al. [24], who suggested using a Boosted Deep Belief Network. This network consisted of several deep networks learning features, and some of these networks get boosted based on their performance. Finally, in [22], Liu et al. introduced CNNs to the problem, with their CNN Ensemble network. The network consisted of three different convolutional networks, which proved to achieve better results than a single CNN.

2.2 Facial Expression Recognition in Videos

When classifying emotion in videos, there is additional information that can be extracted by accounting for the way the face changes over time. There are several papers which attempt to do this, which I will now go over, beginning with aggregation techniques. Aggregation techniques classify each frame within a video, and then combine the results with some sort of aggregator. In [15] and [16], S. Kahou et al. split the video into 10 sections, and within each section aggregate the frame predictions with an average. They go on to use a SVM to classify using all 10 aggregate predictions. An alternative method was introduced in [23] where M. Liu et al. showed that features from individual frames could be mapped to linear subspaces, covariance matrices or Gaussian distributions, allowing these to be passed to a support vector machine.

An alternative approach focussed on attempts to classify the level of emotional intensity present in an individual frame - the idea being that you could derive the emotion of a video base on the strongest emotions present. In [41] X. Zhao et al. propose a network which minimizes the differences between an emotion at low intensity and the same emotion at high intensity, as a way to get better classification of low intensity emotion. The downside of their technique was that it required a training set consisting of pairs of the same emotions at different intensities. To address this, in [4] J. Chen et al. used unsupervised clustering and a semisupervised SVM to detect peak and neutral frames in a large dataset.

Finally, deep spatio-temporal networks were introduced which use sequences of frames as inputs to the networks, in the hopes that we can learn more information from the temporal dynamics of the images. C3Ds [37] are the natural generalization of a CNN to 3 dimensions - instead of convolving an image with a 2D kernel, we convolve a sequence of frames with a 3D kernel. These 3D techniques have been bought to video emotion recognition, for example by X. Ouyang in [27]. An alternative approach was taken in [19] where D. Kim et al. tracked facial landmarks to generate trajectories which were used as features. Finally, a common approach is to use a CNN to learn spatial features of an image followed by a LSTM to learn the temporal features of the overall video, as in [18].

2.3 Reduced resolution classification

There have been a couple of studies looking at how to cope when the face that needs to be classified is at a low resolution. Firstly, Y. Tian [36] looked at how low spatial resolution would affect three steps in a facial expression analysis pipeline: face acquisition, facial data extraction and finally expression recognition. Also, R. Khan et al. produced a framework for expression recognition on low-resolution images [17], by suggesting a new set of features which work on low resolution images - called pyramid of local binary pattern. Finally, T. Vo et al. proposed the pyramid with super resolution architecture [39], which made use of super resolution networks in order to scale up low resolution images with minimal artefacts.

3 METHODOLOGY

In this section, the goal is to build to a classifier which we can use to test video at multiple spatial and temporal resolutions. We first introduce the datasets used in the work, then discuss the image classifiers produced, before finally talking about the techniques used to classify the videos.

3.1 Datasets

For learning facial expressions of images I used the FER+ dataset [1] from E. Barsoum et al., a large collection of images first as part of the FER-2013 dataset [9]. The FER-2013 dataset consists of 36,685 48x48 greyscale images obtained by searching for different emotions on Google Images. The dataset is split into 7 categories - angry, sad, disgust, fear, happy, surprise and neutral. FER+ uses the same collection of images as in FER-2013, but updated the labels to be more accurate by crowd-sourcing the labelling and getting 10 taggers to label each image. In this work, we use majority voting to decide the ground truth of the images, discarding those for which there is no majority.

To learn facial expressions in videos, I used the RAVDESS [25] dataset by SR. Livingstone and FA. Russo. The dataset consists of 24 professional actors speaking two different statements while expressing one of eight emotions. The emotions include calm, happy, sad, angry, fearful, surprise, disgust and neutral. The data is provided at 720p, 30 fps with a blank background, in laboratory settings.

To pre-process the video data, we crop the images to a box around the face. In order to detect the bounding box around the face, we make use of Haar cascades. The purpose of this study is not to focus on how to detect faces at a reduced resolution as there is sufficient work already on that topic [42], [36].

3.2 Training Image Recognition

I used a ResNet50 CNN architecture which is a 50 layer versions in the family of Residual Networks, designed by K. He et al. [12]. They were designed to tackle image recognition tasks, and were significantly deeper than previously introduced architectures, which allowed them to gain accuracy in several recognition tasks. Further, despite the increase in depth, they managed to make the networks relatively easy to optimize.

Instead of training the network beginning with random parameters, I made use of transfer learning - where you use network weights obtained from a different task. This helps to bootstrap the network, giving it a good place to start learning the task you want. We use two different base models, which are publicly available - the first is trained on ImageNet [30], a massive dataset of over a million images, sorted into various nouns. Also, we use weights trained on the VGG Face dataset [28], which consists of 9,131 images of different people's faces. To train, I take the top layers off the pre-trained network, and then add two dense layers to perform the classification on my task. I freeze the original base of the network and train just the new top of the network. After this is complete, I unfreeze the network and train the entirety of the network.

To train the network we use stochastic gradient descent, and in order to obtain good hyperparameters we use Bayesian optimization [33]. Bayesian optimization represents the problem as a Guassian process, which maps hyperparameters to the optimization criteria. By doing this, it can focus the hyperparameter search on those regions which are most likely to lead to good results.

3.3 Training Video Recognition

To classify video, I decided to make use of the best network I previously used for classifying images with the logic that the patterns learnt there would be applicable in the video domain too. In order to extend the network that was designed for single images to video, with several frames, I made use of 3 different techniques. The first two are quite simplistic - I classify each frame of the video before then taking some aggregate of those results. For the first technique, this aggregate is a mean of all the classification scores. The second technique is where I take the max class for each frame, and then take the mode of these classifications.

The third technique I used was a LSTM network, appended to the output of the ResNet, to form a CNN-LSTM architecture. I used a LSTM with 1024 units, which means the LSTM outputs 1024 different features. I follwed the LSTM with two dense layers for the final classifcation. The model was trained using the Adam optimizer [20], which promises to be well suited for optimization problems with a large number of parameters and data. As input to the network, I use 30 frames sequences from the video, and if the video was shorter than 40 frames I repeat frames until we get 30.

To preprocess the video, I created copies of the video at 3 different resolution and 3 different frame rates, for a total of 9 different clips for each input. The resolutions and frame rates were chosen such that they give a wide range of what may be possible over the internet. For resolutions, we had 720p (1280x720 px), 360p (480x360 px) and 144p (256x144 px), and for frame rates we had 30 fps, 15 fps and 5 fps. 720 fps at 30 fps is feasible with a fast internet connection, and low proximity to the server, and 144p at 5 fps would happen with a poor internet connection and/or large distance from a server.

4 EVALUATION

In this section I walk through the results of the experiments, starting with the image recognition task and then moving on to the video recognition tasks.

4.1 Image Recognition

In our image recognition task, I trained a ResNet50 network on the 7-class FER+ dataset, using transfer learning, with two different starting points. One network was pre-trained on the ImageNet dataset, and one on the VGG Face dataset. The recall and F1 score for each class, along with the overall accuracy are plotted in Table 1. The table clearly demonstrates that the VGG pre-train outperforms ImageNet in each class, which is likely due to the fact that VGG model had been trained to recognize specific facial features, whereas ImageNet model is trained on a much wider set of objects so is less dedicated to recognizing faces.

One area where both models underperform is in classifying disgust. This is largely due to the fact that the FER+ dataset is quite unbalanced, with very few disgust expressions within. Furthermore, the line between disgust and anger is quite small, meaning that 47% of the disgust images end up being labelled as anger, as can be seen in the confusion matrix in Figure 2. There are a couple of ways to address this in future, for example I could utilize a weighted loss

Table 1: Classification results on the test set of the FER+ dataset.

model	anger		disgust		fear		happiness		neutral		sadness		surprise		Overall
	Recall	F1	Recall	F1	Recall	F1	Recall	F1	Recall	F1	Recall	F1	Recall	F1	Accuracy
RN50 on VGG	**0.86**	**0.85**	**0.33**	**0.45**	**0.49**	**0.58**	**0.95**	**0.94**	**0.65**	**0.71**	**0.85**	**0.86**	**0.91**	**0.88**	**0.86**
RN50 on ImageNet	0.63	0.67	0	0	0.21	0.34	0.89	0.88	0.44	0.53	0.81	0.8	0.9	0.82	0.78

function which gives more weight to the minority classes. Also I could use under-sampling of majority classes and over-sampling for minority classes. Finally, I could use a Generative Adversarial Network to generate more artificial faces for the minority classes. However, I did not employ these techniques in this work, as my focus was on classifying videos.

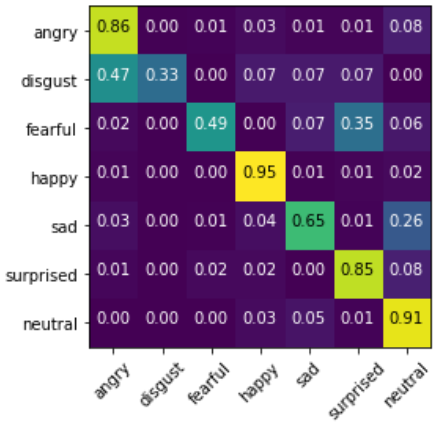


Figure 2: Confusion matrix ResNet50 pre-trained on VGG Face Dataset.

4.2 Video Recognition

In the video recognition task, I tested three main ways in which we could use the ResNet50 model in order to predict the sentiment in videos. The dataset I used was the 8-class RAVDESS emotional video dataset. Here I have plotted the confusion matrix for the best results (over each fps and resolution) for each method in Figure 3. Additionally, in Table 2 I have plotted the accuracy of each method for every resolution combination - note accuracy is a useful metric here as the RAVDESS dataset is balanced. In this section, I will first discuss the performance of the mean and mode classifier, discussing their performance with different emotion classes before talking about their performance with the varied fps and resolution. I will then do the same for the LSTM classifier and finally talk about some interesting anomalies in the results.

The mean and mode classification methods both performed poorly, which is unsurprising as they both could not use the temporal dynamics, such as people's facial movements in order to determine emotion. As can be seen in the confusion matrices in Figures 4a and 4b, with the exception of happiness most of the emotions end up being classified as neutral. This is likely as the RAVDESS dataset consists of people in speech, and while speaking it is hard to tell what emotion a person is displaying, by just looking at a frame. These methods would definitely not perform well in a cloud robotics scenario, as the majority of frames are likely to be neutral when a person is interacting with the robot, and these methods are unable to differentiate between a scenario where the emotion is actually neutral, and scenarios where for most of the frames the expression is unclear.

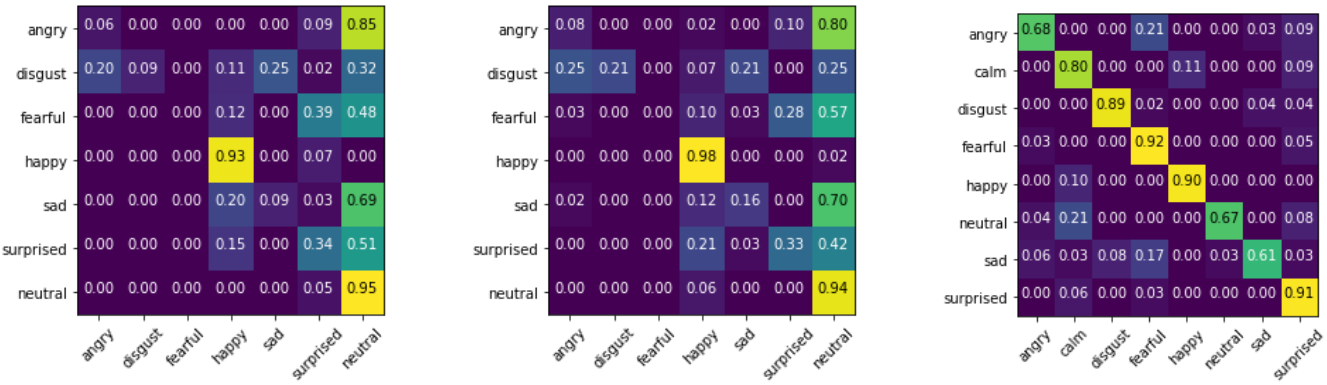
For mean and mode classification methods, there are some minor differences across different framerates and resolutions, but they are likely not significant. The differences are most likely due to each different framerate getting a slightly different distribution of frames, meaning lower framerates might miss important frames, but equally, higher framerates may get many more neutral frames.

The LSTM method performed very well, achieving a best accuracy of 81%, compared to mean and mode which both managed a maximum of 33%. It is clear that the LSTM was able to learn the patterns in the changes in temporal dynamics that represent different emotions. By looking at the confusion matrix in Figure 4c, it is clear to see that the classifier performs well across all emotions, with a minimum recall of 61%, for the sad class.

Additionally, the LSTM performs well across all resolutions and framerates, with most of the results scoring an accuracy of 75% or higher. Surprisingly, the one exception to that is the highest resolution/framerate results, which scored only 67%, as I will discuss later. I was surprised to find that the performance does not degrade significantly as framerate decreases, as I had previously assumed that as the LSTM had learnt temporal dynamics at 30 fps, changing framerate would change those dynamics. The fact that this classifier can perform very impressive results with 5 fps and 256x144 p resolution implies it would perform very well in a cloud robotics scenario, where the video data may come in at reduced resolution due to huge variance in network conditions and distance from cloud endpoints.

One interesting artefact in the data is the performance at the highest resolution (1280x720 px and 30 fps) - across all three classification methods this combination resulted in the lowest accuracy. There are a couple of reasons that this could be the case, firstly due to the choice of image dataset. The FER+ dataset consists of 48x48 px images of faces, which is comparatively low resolution. As a result of this, it is possible that the CNN architecture is most familiar with lower resolution images of faces. This is not necessarily a negative though, as it results in the consequence that the model performs very well on low resolution images, which was the aim of this study. An easy fix for this problem would be to downsample the images before classifying. Alternatively, a dataset with higher resolution could be used to train the CNN, however this may have the undesired effect of degrading the network performance on low resolution data. In the case of the LSTM classifier, it is especially surprising, as the LSTM was trained on 30 fps video. I am not sure why it performed so poorly at this framerate, and would

Figure 3: Confusion matrices for video classification on the RAVDESS dataset, based on the best results for any fps/resolution.



(a) Mean Classification **(b) Mode Classification** **(c) LSTM Classification**

Table 2: Comparison of video classifiers accuracy on the RAVDESS dataset.

(a) Mean Classification

	Framerate (fps)		
Resolution (px)	30	15	5
1280x720	0.26	0.31	0.30
640x360	**0.33**	0.30	0.28
256x144	**0.33**	0.29	0.31

(b) Mode Classification

	Framerate (fps)		
Resolution (px)	30	15	5
1280x720	0.27	0.31	0.31
640x360	0.32	0.30	0.30
256x144	**0.33**	0.31	0.32

(c) LSTM Classification

	Framerate (fps)		
Resolution (px)	30	15	5
1280x720	0.67	**0.81**	0.77
640x360	0.77	0.77	0.75
256x144	0.80	**0.81**	0.76

like to investigate further, however for now an easy fix would be to downsample the frames, and only accept frames at 15 fps.

5 FUTURE WORK

In this section I go over the directions this work could take in the future. Firstly, an important next step would take a look at the effect of varying resolutions on multimodal models, where we don't just consider visual inputs to the network, and instead use cues from the audio and even textual modalities. Multimodal models are important as it is not always possible to derive a prediction from just one modality. It will be interesting to see how resolution affects the variety of modalities, for example seeing if changing the sample rate and/or bit depth of the audio has an impact on audio classification. There are several interesting approaches that could be taken, for example to prepare for ultra-low bandwidth connection, would it be possible to build a model that dynamically chooses what data should be sent over the network. Perhaps it would send both audio and visual data when network connection is good, but only send audio data when the network connection fades. Furthermore, it may be possible to do some basic classification on the client-side - in choosing which frames to send to the server for further analysis. If the client was able to pick frames where emotions are more clear then it would be able to send much fewer frames and use less bandwidth.

Another direction this work could go in to is looking at in-the-wild datasets, where data is collected from real-world scenarios instead of in the lab, as in this study. This is an important step as when the robots are entered into the real world it is vital that they are not constrained to lab conditions, and can classify data in real-world conditions, fostering beneficial interactions wherever they might need to be. The image model I trained is already prepared for in-the-wild, as the FER+ dataset is based off of images taken from a wide range of sources. However, the video LSTM model may not perform well on an in-the-wild dataset, as it was trained in lab conditions - this would need to be tested. Additionally, in-the-wild conditions may result in especially low resolution images - if a person is far from a camera then there face would be especially low resolution. It may be worth investigating whether some pre-processing could be done on the client, in order to crop to around the individual's face before transmitting over the network.

An additional future possibility would be to look at other advanced neural network architectures which have been previously used for classification. For example, T. Vo et al.'s pyramid super-resolution architecture [39] which uses super-resolution networks to reliably scale up facial images and then classify on those images. It has shown to be effective on in-the-wild image datasets, so would be interesting to test on in-the-wild videos, by finding a way to append a LSTM to the architecture. It may also be useful to test other networks instead of the LSTM, for example transformers [38] which make use of attention layers, that allow models to be trained in less time, and often achieve superior results to LSTMs.

Another future direction may looks at reducing the size of the models, as currently the CNN-LSTM model is very large, with 24,693,320 parameters. By reducing the size of the network, it is possible to classify more frames, use less energy and save money.

In order to reduce the size of the models, there are several possible techniques that could be employed, the first of which is training a smaller model - I could look at using ResNet18 or another lighter-weight network, followed by a LSTM with fewer states, in an attempt to find a good balance between accuracy and model size. An alternative approach would be pruning - where we find weights or layers that have a low importance to the result, and remove them from the network, which can sometimes even result in improved accuracy.

Finally, it is important than someone looks into how this technique could be implemented in a social robotics scenario. There are several possible directions this could be taken, firstly it could be worthwhile to produce a model that can determine a confidence bound over it's predictions. The confidence bounds may vary based on the available resolutions, and how confident it is about the data. This would allow the robot to react cautiously if the emotion classification is uncertain. Additionally, it may be worth investigating whether a very light-weight network could be run on the robot itself. This would allow the robot to perform some offline calculations in case the network is unavailable, and the robots could switch between an online and offline mode to to avoid interruptions caused by network disconnections. Finally, some studies could be done to see if human subjects can perceive the difference in emotional response between a robot which performs sentiment analysis on the cloud and one which has dedicated hardware to perform classification.

6 CONCLUSION

This work applied different techniques to classify sentiment in video at various spacial and temporal resolutions, to test the techniques adaptability for use in a cloud processing scenario. The findings show that with a CNN-LSTM model very good performance can be achieved, up to 81% classification accuracy with 8-class classification on the RAVDESS dataset. Additionally, results show that this technique can still produce 76% accuracy on very low spatial and visual resolution video, suggesting that it could be utilized in a cloud processing framework.

REFERENCES

[1] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. 2016. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. 279–283.

[2] Marian Stewart Bartlett, Gwen Littlewort, Mark Frank, Claudia Lainscsek, Ian Fasel, and Javier Movellan. 2005. Recognizing facial expression: machine learning and application to spontaneous behavior. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 2. IEEE, 568–573.

[3] Oya Celiktutan, Evangelos Sariyanidi, and Hatice Gunes. 2018. Computational Analysis of Affect, Personality, and Engagement in Human–Robot Interactions. In *Computer Vision for Assistive Healthcare*. Elsevier, 283–318.

[4] Jingying Chen, Ruyi Xu, and Leyuan Liu. 2018. Deep peak-neutral difference feature for facial expression recognition. *Multimedia Tools and Applications* 77, 22 (2018), 29871–29887.

[5] Diligent Robotics. 2019. Moxi — Diligent Robotics. https://www.diligentrobots.com/moxihttps://diligentrobots.com/moxi

[6] P Ekman, P.E.E.L. Rosenberg, P H D of Psychology Paul Ekman, E L Rosenberg, L.D.P.E.L. Rosenberg, and M B Smith. 1997. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford University Press. https://books.google.co.uk/books?id=fFGYs079-7YC

[7] Elliq. 2019. ElliQ, the sidekick for happier aging – Intuition Robotics. https://elliq.com/

[8] E Friesen and Paul Ekman. 1978. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto* 3, 2 (1978), 5.

[9] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, and Others. 2013. Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing*. Springer, 117–124.

[10] Alex Graves, Santiago Fernández, Marcus Liwicki, Horst Bunke, and Jürgen Schmidhuber. 2008. Unconstrained online handwriting recognition with recurrent neural networks. In *Advances in Neural Information Processing Systems 20, NIPS 2008*.

[11] Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. In *Neural Networks*, Vol. 18. Pergamon, 602–610. https://doi.org/10.1016/j.neunet.2005.06.042

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *CoRR* abs/1512.0 (2015). arXiv:1512.03385 http://arxiv.org/abs/1512.03385

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2016-Decem. IEEE Computer Society, 770–778. https://doi.org/10.1109/CVPR.2016.90 arXiv:1512.03385

[14] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[15] Samira Ebrahimi Kahou, Xavier Bouthillier, Pascal Lamblin, Caglar Gulcehre, Vincent Michalski, Kishore Konda, Sébastien Jean, Pierre Froumenty, Yann Dauphin, Nicolas Boulanger-Lewandowski, and Others. 2016. Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces* 10, 2 (2016), 99–111.

[16] Samira Ebrahimi Kahou, Christopher Pal, Xavier Bouthillier, Pierre Froumenty, Çaglar Gülçehre, Roland Memisevic, Pascal Vincent, Aaron Courville, Yoshua Bengio, Raul Chandias Ferrari, and Others. 2013. Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International conference on multimodal interaction*. 543–550.

[17] Rizwan Ahmed Khan, Alexandre Meyer, Hubert Konik, and Saida Bouakaz. 2013. Framework for reliable, real-time facial expression recognition for low resolution images. *Pattern Recognition Letters* 34, 10 (2013), 1159–1168.

[18] Dae Hoe Kim, Wissam J Baddar, Jinhyeok Jang, and Yong Man Ro. 2017. Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. *IEEE Transactions on Affective Computing* 10, 2 (2017), 223–236.

[19] Dae Ha Kim, Min Kyu Lee, Dong Yoon Choi, and Byung Cheol Song. 2017. Multimodal emotion recognition using semi-supervised learning and multiple neural networks in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. 529–535.

[20] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[21] Steve Lawrence, C Lee Giles, Ah Chung Tsoi, and Andrew D Back. 1997. Face recognition: A convolutional neural-network approach. *IEEE transactions on neural networks* 8, 1 (1997), 98–113.

[22] Kuang Liu, Mingmin Zhang, and Zhigeng Pan. 2016. Facial expression recognition with CNN ensemble. In *2016 international conference on cyberworlds (CW)*. IEEE, 163–166.

[23] Mengyi Liu, Ruiping Wang, Shaoxin Li, Shiguang Shan, Zhiwu Huang, and Xilin Chen. 2014. Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. In *Proceedings of the 16th International Conference on multimodal interaction*. 494–501.

[24] Ping Liu, Shizhong Han, Zibo Meng, and Yan Tong. 2014. Facial expression recognition via a boosted deep belief network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1805–1812.

[25] Steven R Livingstone and Frank 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* 13, 5 (2018), e0196391.

[26] No Isolation. 2018. AV1 – the robot for children with long-term illness. https://www.noisolation.com/global/av1/https://www.noisolation.com/global/av1/privacy-and-resources/{%}0Ahttps://www.noisolation.com/global/av1/

[27] Xi Ouyang, Shigenori Kawaai, Ester Gue Hua Goh, Shengmei Shen, Wan Ding, Huaiping Ming, and Dong-Yan Huang. 2017. Audio-visual emotion recognition using deep transfer learning and multiple temporal models. In *Proceedings of the 19th ACM international conference on multimodal interaction*. 577–582.

[28] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep face recognition. (2015).

[29] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1985. *Learning internal representations by error propagation*. Technical Report. California Univ San Diego La Jolla Inst for Cognitive Science.

[30] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge.

International Journal of Computer Vision (IJCV) 115, 3 (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

[31] Md. Saiful Islam and Emam Hossain. 2020. Foreign Exchange Currency Rate Prediction using a GRU-LSTM Hybrid Network. *Soft Computing Letters* (oct 2020), 100009. https://doi.org/10.1016/j.socl.2020.100009

[32] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR. arXiv:1409.1556 http://www.robots.ox.ac.uk/

[33] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. 2012. Practical bayesian optimization of machine learning algorithms. *arXiv preprint arXiv:1206.2944* (2012).

[34] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, A Rabinovich, and Others. 2014. Going deeper with convolutions. arXiv 2014. *arXiv preprint arXiv:1409.4842* 1409 (2014).

[35] Y-I Tian, Takeo Kanade, and Jeffrey F Cohn. 2001. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence* 23, 2 (2001), 97–115.

[36] Ying-li Tian. 2004. Evaluation of face resolution for expression analysis. In *2004 Conference on Computer Vision and Pattern Recognition Workshop*. IEEE, 82.

[37] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762* (2017).

[39] Thanh-Hung Vo, Guee-Sang Lee, Hyung-Jeong Yang, and Soo-Hyung Kim. 2020. Pyramid with Super Resolution for In-the-Wild Facial Expression Recognition. *IEEE Access* 8 (2020), 131988–132001.

[40] Jacob Whitehill and Christian W Omlin. 2006. Haar features for FACS AU recognition. In *7th international conference on automatic face and gesture recognition (FGR06)*. IEEE, 5—-pp.

[41] Xiangyun Zhao, Xiaodan Liang, Luoqi Liu, Teng Li, Yugang Han, Nuno Vasconcelos, and Shuicheng Yan. 2016. Peak-piloted deep network for facial expression recognition. In *European conference on computer vision*. Springer, 425–442.

[42] Jun Zheng, Geovany A Ramirez, and Olac Fuentes. 2010. Face detection in low-resolution color images. In *International Conference Image Analysis and Recognition*. Springer, 454–463.