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 (cloud robotics), 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. We 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 TODO 7-class problem, we can achieve 73% accuracy on high resolution video, and maintain 66% accuracy with the lowest resolutions. To the best of our 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, 3 pages. https://doi.org/10.1145/1122445.1122456

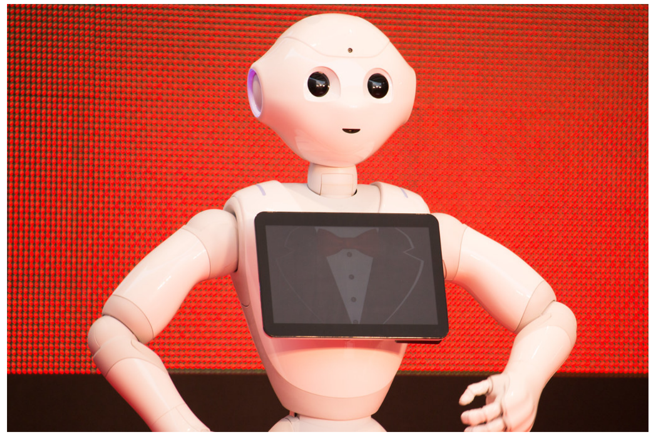

Figure 1: Pepper by SoftBank Robotics, an example of a social robotic platform which has been trialled in care homes in the UK. Photograph by Dick Thomas, via creative commons. (source)

1 INTRODUCTION

Social robots are becoming increasingly widespread, with uses in a wide range of locations, providing help in hospitals [4], care homes [6] and schools [19]. 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 [2]. 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.

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 [15] and object detection and classification [25]. Several architectures have been proposed offering impressive ability to learn features from video using purposes, for example the VGG16 network [24] and the ResNet50 network [10], both of which were able to achieve winning results in the ImageNet object detection and classification challenge [22].

For classifying videos, it is often vital to take temporal data into account, and as a result Recurrent Neural Networks (RNNs) [21] 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 [11] (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 [9], market prediction [23] and handwriting recognition [8].

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

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

are tested on a 7-class video dataset, and our results show that we can achieve an accuracy of X on

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 will give an overview of facial expression recognition techniques, first for images, and then for videos, followed by a section going over progress on classifying data with reduced resolution.

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) [7], 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 [5]. 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 [26]. 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) [1]. 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 [28].

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. [18], 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. In [16], 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. Finally, in TODO

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 [12] and [13], 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 [17] 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 [29] 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 [3] 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 [27] 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 [20]. An alternative approach was taken in [14] 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 [**?**].

3 FACIAL EXPRESSION RECOGNITION

3.1 Images

4 METHODOLOGY

4.1 Transfer Learning

In order to efficiently train on

5 FUTURE WORK

6 CONCLUSION

ACKNOWLEDGMENTS

To Robert, for the bagels and explaining CMYK and color spaces.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

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

[5] 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

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

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

[8] 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*.

[9] 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

[10] 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

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

[12] 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.

[13] 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.

[14] Dae Ha Kim, Min Kyu Lee, Dong Yoon Choi, and Byung Cheol Song. 2017. Multi-modal 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.

[15] 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.

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

[17] 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.

[18] 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.

[19] 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/

[20] 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.

[21] 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.

[22] 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

[23] 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

[24] 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/

[25] 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).

[26] 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.

[27] 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.

[28] 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.

[29] 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.