Visual Saliency Maps Can Apply to Facial Expression Recognition

Zhenyue Qin\textsuperscript{1}, Jie Wu\textsuperscript{2}, Yang Liu\textsuperscript{1,3} and Tom Gedeon\textsuperscript{1}

\textsuperscript{1}Australian National University, Canberra, Australia
\textsuperscript{2}Sun Yat-sen University, Guangzhou, China
\textsuperscript{3}DATA61, CSIRO, Canberra, Australia

Abstract

Human eyes concentrate different facial regions during distinct cognitive activities. We study utilising facial visual saliency maps to classify different facial expressions into different emotions. Our results show that our novel method of merely using facial saliency maps can achieve a descent accuracy of 65\%, much higher than the chance level of $1/7$. Furthermore, our approach is of semi-supervision, i.e., our facial saliency maps are generated from a general saliency prediction algorithm that is not explicitly designed for face images. We also discovered that the classification accuracies of each emotional class using saliency maps demonstrate a strong positive correlation with the accuracies produced by face images. Our work implies that humans may look at different facial areas in order to perceive different emotions.

1. Introduction

Studies in visual cognition show that humans do not focus on each part of a given image the same during observing a scene \cite{31}. Visual saliency algorithms in computer vision aims to determining the eye-catching objects that are consistent with human perception \cite{38}. This saliency detection is essential for many applications in graphics, design, and human computer interaction \cite{18}. For instance, saliency prediction can localise the most eye-catching parts of an image so it can guide video compression. It can also be applied in image captioning \cite{57}, advertisement design and so on.

Traditional saliency prediction algorithms, prior to the deep learning revolution, leverage a range of low-level handcrafted clues at multiple scales, such as colours and textures, and combine them to form a saliency map \cite{14}.

Conventional machine learning techniques have also been applied to detect saliency regions such as the Markov chain \cite{18} and regression models \cite{16}. Recently, with the advent of deep learning, the state of the art accuracy of saliency detection have gained a dramatic improvement \cite{6}. Moreover, the new proposed attention-based mechanism can help deep learning models deal with cluster and scale up to large input images \cite{50}.

Saliency prediction can shed light in the research of cognition interpretation and facial expression recognition. There have been extensive studies on the relationship between cognitive activities and eye fixation targets \cite{19, 20}. For example, previous studies reveal that the concentrating regions of people’s eyes on their opposites’ faces vary given different latent cognitive activities \cite{29}. Furthermore, people’s eye fixation clusters will differ during telling the truth (i.e., I know the person) and deceiving (i.e., I don’t know the person) given a familiar face image \cite{59}, as Figure 1 shows.

\begin{figure}[h]
\centering
\includegraphics[width=0.4\textwidth]{Fig1.png}
\caption{Eye fixation clusters are different during telling the truth (I know the person) and deceiving (I don’t know the person) given a familiar face image. A lot of the fixation points of telling the truth locate on the eye regions, whilst almost none for lying.}
\end{figure}

One remarkable capability that differs humans from other creatures is to accurately and efficiently denote emotional expressions from others’ faces \cite{32}. These facial ex-
pressions provide clues about the person’s cognitive loads, intention, and emotional state [9]. They can even convey more information than normal communication methods like utterance [12]. Therefore, computer vision systems that are capable of understanding human emotions from face images have been intensely studied for decades to facilitate computers collaborate with humans better [8].

In this paper, we investigate the relationship between saliency prediction on face images and facial expression recognition. A human face image of happiness and its corresponding saliency map is illustrated in Figure 2. To the best of our knowledge, we are the first to explore the possibility of merely utilising saliency prediction outcomes generated from the state-of-art deep-learning-based saliency detection algorithms to discern facial expressions of different emotions. In summary, our main contributions are three-fold:

1) Our problem domain is novel. We are the first to raise the possibility of recognising facial expressions by merely using saliency predicted results. The conclusions of this study can be applied to understanding the process of humans perceiving emotions, designing virtual faces that are more emotionally real, being a test bed for evaluating saliency prediction algorithms, and so on.

2) We have achieved a descent accuracy for a seven-class classification problem (50% on FER-2013 dataset [13] and 63% on CK+ dataset [22]), which is above the expected random classification result (1/7) with a large margin.

3) Our approach is semi-supervised. We did not explicitly utilise the facial images to train the saliency detection algorithm. Instead, we applied a pre-trained saliency prediction model that is generated from the datasets of SALICON [17], MIT1003 and CAT2000 [3] to produce saliency maps. These two datasets include a variety of image classes other than human faces, where Figure 3 presents a subset of them.

Furthermore, Our results confirm that human eyes fixate on different facial regions for distinct emotional expressions [23].

2. Related Work

In this section, we briefly introduce the developments of facial expression recognition and saliency prediction, and review the current work on utilising saliency prediction techniques for recognising facial expressions.

2.1. Facial Expression Recognition

The Facial Action Coding System (FACS), is one of the early pioneering work on facial expression recognition [10]. Its proposers, Ekman et al., made an in-depth study on the mechanisms of facial muscle movements and how these muscles control the emotional expressions [11]. They further divided the human face into 46 independent and interconnected action units (AU), and mapped each unit with its corresponding emotions [10]. Figure 4 illustrates a concrete example of FACS. Although this approach is easy and intuitive, it is not practical since it requires a laborious amount of work by experts to manually label the types of facial muscle movements.

Subsequently, researchers started applying conventional machine learning techniques. These methods include Principal Component Analysis [4], Linear Discriminant Analysis [27], and also ensemble methods that compose a range of machine learning algorithms to make a decision. Recently, deep learning based methodologies have significantly enhanced the performance of accurately classifying facial ex-
expressions. Unlike traditional machine learning approaches, which heavily rely on complex rules, deep learning methods generally are simple and end-to-end. That is, they can learn the representative features of different facial expressions spontaneously.

2.2. Saliency Prediction

Similar to the development of facial expression recognition, only recently, saliency prediction has achieved a significant enhancement due to the emergence of deep learning. Unlike traditional saliency prediction algorithms, which greatly depend on the quality of handcrafted features and complex human-made rules, deep learning saliency detection methods can automatically learn useful features during training through techniques like Convolutional Neural Networks (CNNs). Even more recently, the attention mechanism has demonstrated its superior performance in Natural Language Processing (NLP) tasks, such as machine translation and grammar correction. Therefore, various works have also explored applying the attention model on visual tasks and achieved descent results as well.

2.3. Saliency-Based Facial Expression Recognition

To the best of our knowledge, the two very recent publications that explore using saliency predicted results to increase the performance of facial expression recognition. We summarise our novelties compared with the above two works as following:

1) Both use detected saliency maps as additional features in order to make better classifications. In contrast, we utilise the generated saliency maps as the only inputs. Our attempt is novel.

2) Unlike only utilising traditional statistical algorithms for saliency prediction and facial expression recognition, which are complex and consider an extensive of features other than saliency maps, we apply the state-of-art deep learning based methods to tackle both of these two tasks. Our methods are simple and intuitive.

3. Our Framework

Our proposed method consists of two components, namely the saliency map generator and the facial expression classifier. We apply the state-of-art deep learning based saliency predictor with visual attention mechanisms to produce our saliency maps. Its overall structure is given in Figure 5. For more information, please refer to 6.

As to our facial expression classifier, we apply a deep convolutional neural network for an end-to-end feature extraction and decision making. We consider both ResNet-18 and VGG-19 with slight customisation, specifically:

1) We introduced the dropout mechanism between the final convolutional layer and the fully-connected layer.

2) Instead of using several fully-connected layers in the end, we only add one fully connected layer, followed by another fully-connected layer with softmax to classify the input into one of seven classes.

We apply VGG-19 in our experiments since it performs better than ResNet-18 in our case. Furthermore, we use cross-entropy as our loss function to correspond our choice of softmax in the final layer. This choice is because cross-entropy can to extent tackle noise labels, and using the cross-entropy error function can lead to a faster training as well as improved generalisation than sum-of-squares. Figure 5 illustrates the architecture of our VGG-19 model.

4. Dataset

We evaluated performance of our proposed methodology on two different common-applied datasets in order to test generality, namely FER2013 and CK+. We give brief explanations about these two datasets.

4.1. FER2013

The FER2013 dataset comes from a challenge in representation learning, specifically, facial expression recognition challenge held on Kaggle. It consists of 48 × 48 pixel grayscale images of faces, and has 28709 training images, 3589 public test images and another 3589 examples for private testing. These facial expression images are labelled with seven emotional classes, i.e., angry, disgust, fear, happy, sad, surprise, neutral. Figure 7 presents examples of each expression class from the FER2013 dataset. This classification is based on Ekman et al.’s early pioneering studies on facial expression recognition, which re-
ports the above seven emotions are universally recognisable with respect to different cultural backgrounds. Furthermore, these face images were sourced from the Internet with noises and relatively low quality. Figure 8 gives examples of noises within the FER2013 dataset. However, we did not conduct data cleansing for the fare purpose.

![Figure 5: Overview of the saliency prediction model [6].](image)

![Figure 6: The architecture of our customised VGG-19 model.](image)

![Figure 7: Examples of FER2013 dataset.](image)

(a) Angry  (b) Disgust  (c) Fear  (d) Happy  (e) Sad  (f) Surprise  (g) Neutral

![Figure 8: Noise examples of FER2013 dataset. (a) an face image with watermark; (b)&(c) random pictures.](image)

4.2. CK+

The Extended Cohn-Kanade (CK+) database was initially released in 2010 as an extension of the original Cohn-Kanade (CK) database, which was published in 2000 [26]. It consists seven emotional classes, namely angry, disgust, fear, happy, sad, surprise, contempt. That is, compared with the FER2013 dataset, it replaces neutral with contempt. The facial expressions of this dataset were obtained under laboratory-controlled environments, therefore are relatively more rigorous and reliable compared with the FER2013
dataset. The CK+ dataset is widely used in expression-related studies \cite{26}. Figure \ref{fig:ckplus_examples} gives examples corresponding different facial expressions in the CK+ dataset.

![Figure 9: Examples of CK+ dataset.](image)

(a) Angry       (b) Disgust       (c) Fear        (d) Happy

![Figure 9: Examples of CK+ dataset.](image)

(e) Sad         (f) Surprise      (g) Contempt

5. Experimental Results

5.1. Expression Classification With Face Images

We first fine-tuned our facial expression classifier in order to maximise the potential performance of the utilised deep convolutional neural network. In order to prevent overfitting and increase prediction robustness, we conducted data augmentation to the used dataset. Specifically, we randomly created 10 cropped image of size $44 \times 44$ for each the original image, whose size is $48 \times 48$. Furthermore, we also collected 10 processed images for each facial expression to test by cropping the upper left corner, the lower left corner, the upper right corner, the lower right corner, the centre, and subsequently taking the reflection of each of these cropped images. We make the final decision by taking the average results of these 10 processed images to reduce the classification error.

We have achieved an accuracy of 71% on the FER2013 public test and 73% accuracy on its private test. Our learning rate is 0.01 and the FER2013 and CK+ dataset epoch number are 250 and 60, respectively. For more detailed hyper-parameter information and source code, please refer \cite{36}.

Table \ref{tab:comparison} gives a performance comparison on the FER2013 private test dataset between our facial expression classification approach with state-of-art approaches.

\begin{table}[h]
\centering
\begin{tabular}{|l|l|}
\hline
\textbf{Approach} & \textbf{Accuracy} \\
\hline
Best FER2013 Record \cite{21} & 71\% \\
DNNRL \cite{2016} \cite{24} & 71\% \\
CPC \cite{2018} \cite{5} & 73\% \\
\textbf{Ours} & \textbf{73}\% \\
\hline
\end{tabular}
\caption{The comparison between our facial expression classification approach with state-of-art approaches on the FER2013 private test dataset.}
\end{table}

5.2. Expression Classification Using Saliency Maps

We tested the classification accuracy of merely utilising facial saliency maps generated by \cite{6} to discern the different facial expressions of different emotions. Table \ref{tab:saliency_results} presents our accuracies, which are far above the random 1/7 accuracy. We have obtained relatively descent results for both of the datasets also indicate a generality of our approach. Furthermore, this implies that humans may focus on different facial regions in order to perceive distinct emotions of others.

\begin{table}[h]
\centering
\begin{tabular}{|l|l|}
\hline
\textbf{Dataset} & \textbf{Accuracy} \\
\hline
FER2013 Public Test & 48\% \\
FER2013 Private Test & 50\% \\
CK+ (10 Fold) & 63\% \\
\hline
\end{tabular}
\caption{The facial expression classification results generated by merely using facial saliency maps.}
\end{table}

We also obtained the confusion matrices of our facial expression classification using saliency maps. As Figure \ref{fig:confusion_1} \ref{fig:confusion_2} and \ref{fig:confusion_3} illustrate, facial saliency maps did not perform evenly well for all the facial expressions. For example, happiness and surprise are the two most distinguishable emotions, whereas the contempt emotion in the CK+ dataset is almost not recognisable. We hypothesise that it was due to the classification error.

These confusion maps may provide us with insights on where on face do people look at in order to perceive emotions. For instance, we hypothesise it is due to contempt not being one of the six basic emotion and people perceive it by staring to the same facial regions as disgust and happiness.

5.3. Classification Accuracy Correlations Between Using Saliency Maps and Face Images

We obtained the confusion matrices for the facial expression classification accuracies of using the original face images as Figure \ref{fig:confusion_4} \ref{fig:confusion_5} and \ref{fig:confusion_6} demonstrate. We evaluated the Pearson correlation coefficients of the diagonal accuracies generated by using saliency maps and original face images.
Figure 10: The confusion matrix of classifying facial expressions by merely using saliency maps from the FER2013 public test dataset.

Figure 11: The confusion matrix of classifying facial expressions by merely using saliency maps from the FER2013 private test dataset.

Figure 12: The confusion matrix of classifying facial expressions by merely using saliency maps from the CK+ dataset.

Figure 13: The confusion matrix of classifying facial expressions by using face images from the FER2013 public test dataset.

Table 3: Pearson correlation coefficients of accuracies of confusion matrix diagonals from using saliency maps and face images.

| Dataset            | Pearson Correlation Coefficient |
|--------------------|---------------------------------|
| FER2013 Public Test| 0.9277                          |
| FER2013 Private Test| 0.9450                          |
| CK+                | 0.8080                          |

6. Conclusion

In this study, we showed a novel approach that facial merely using saliency maps generated from general visual saliency prediction algorithms can present descent classification accuracies of facial expressions, much higher than the chance level of 1/7. We discovered that the accuracies of each emotion class generated by saliency maps demon-
stratified a strong positive correlation with the results generated by using original facial images. These findings imply that people may perceive different emotions by observing different facial regions. In the future, we aim to determine a specific map between the types of the emotions and the visual concentration regions on human faces. We outlook to embed these mapping outcomes into the Facial Action Coding System (FACS) to enhance the understanding of emotion perception.

References

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[2] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[3] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba. Mit saliency benchmark.

[4] A. J. Calder, A. M. Burton, P. Miller, A. W. Young, and S. Akamatsu. A principal component analysis of facial expressions. *Vision research*, 41(9):1179–1208, 2001.

[5] T. Chang, G. Wen, Y. Yu, and J. Ma. Facial expression recognition based on complexity perception classification algorithm. *arXiv preprint arXiv:1803.00185*, 2018.

[6] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. Sam: Pushing the limits of saliency prediction models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1890–1892, 2018.

[7] P. Ekman. Facial action coding system. [https://www.paulekman.com/product-category/facs/](https://www.paulekman.com/product-category/facs/) Accessed: 2018-11-10.

[8] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

[9] P. Ekman, W. V. Friesen, and P. Ellsworth. *Emotion in the human face: Guidelines for research and an integration of findings*. Elsevier, 2013.

[10] P. Ekman and E. L. Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[11] I. A. Essa and A. P. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 19(7):757–763, 1997.

[12] J.-M. Fernández-Dols, H. Wallbott, and F. Sanchez. Emotion category accessibility and the decoding of emotion from facial expression and context. *Journal of Nonverbal Behavior*, 15(2):107–123, 1991.

[13] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*, pages 117–124. Springer, 2013.

[14] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in neural information processing systems*, pages 545–552, 2007.

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

[16] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li. Salient object detection: A discriminative regional feature integration approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2083–2090, 2013.

[17] M. Jiang, S. Huang, J. Duan, and Q. Zhao. Salicon: Saliency in context. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.
[18] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *Computer Vision, 2009 IEEE 12th international conference on*, pages 2106–2113. IEEE, 2009.

[19] M. A. Just and P. A. Carpenter. Eye fixations and cognitive processes. *Cognitive psychology*, 8(4):441–480, 1976.

[20] M. A. Just and P. A. Carpenter. Cognitive processes in reading: Models based on readers eye fixations. In *Interactive Processes in Reading*, pages 177–213. Routledge, 2017.

[21] Kaggle. *Fer 2013 dataset*. [Accessed: 2018-11-10].

[22] T. Kanade, Y. Tian, and J. F. Cohn. Comprehensive database for facial expression analysis. In *fg*, page 46. IEEE, 2000.

[23] R. A. Khan, A. Meyer, H. Konik, and S. Bouakaz. Saliency-based framework for facial expression recognition. *Frontiers of Computer Science*, pages 1–16, 2017.

[24] B.-K. Kim, J. Roh, S.-Y. Dong, and S.-Y. Lee. Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. *Journal on Multimodal User Interfaces*, 10(2):173–189, 2016.

[25] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[26] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 94–101. June 2010.

[27] M. J. Lyons, J. Budynek, and S. Akamatsu. Automatic classification of single facial images. *IEEE transactions on pattern analysis and machine intelligence*, 21(12):1357–1362, 1999.

[28] V. Mavani, S. Raman, and K. P. Miyapuram. Facial expression recognition using visual saliency and deep learning. In *Computer Vision Workshop (ICCVW), 2017 IEEE International Conference on*, pages 2783–2788. IEEE, 2017.

[29] A. E. Millen, L. Hope, A. P. Hillstrom, and A. Vrij. Tracking the truth: the effect of face familiarity on eye fixations during deception. *The Quarterly Journal of Experimental Psychology*, 70(5):930–943, 2017.

[30] V. Mnih, N. Heess, A. Graves, et al. Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212, 2014.

[31] R. A. Rensink. The dynamic representation of scenes. *Visual cognition*, 7(1-3):17–42, 2000.

[32] P. G. Schyns, L. S. Petro, and M. L. Smith. Transmission of facial expressions of emotion co-evolved with their efficient decoding in the brain: behavioral and brain evidence. *PloS one*, 4(5):e5625, 2009.

[33] P. Y. Simard, D. Steinkraus, and J. C. Platt. Best practices for convolutional neural networks applied to visual document analysis. In *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, pages 958–963, Aug 2003.

[34] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[35] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[36] J. Wu. Facial expression recognition source code. [Accessed: 2018-11-10].

[37] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[38] J. Zhang, T. Zhang, Y. Dai, M. Harandi, and R. Hartley. Deep unsupervised saliency detection: A multiple noisy labeling perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9029–9038, 2018.

[39] J. Zuo, T. Gedeon, and Z. Qin. Your eyes say you’re lying: An eye movement pattern analysis for face familiarity and deceptive cognition, 2018.