User state detection using facial images with mask cover

Danushka Bandara[1][0000-0002-8885-622X]

[1] Fairfield University, Fairfield CT 06824, USA
dbandara@fairfield.edu

Abstract.
Widespread use of masks was mandated in many countries as a direct result of the covid-19 pandemic. This meant that mask wearing, which was previously restricted to specialized occupations or cities with high levels of pollution became the norm in many places of the world. This has obvious implications for any system that uses facial images to infer user state. This work attempts to gauge the effect of mask wearing on such systems. Arousal classification is used in this study due to its well-studied nature in image processing literature. Using "Affect in the wild" video dataset, the "masks" were synthetically placed on the facial images extracted from videos. A binary classification between high and low arousal shows that there is a drop in accuracy when using masks. However, this drop is larger in across subject classification than within subject classification. The study shows that it is feasible to develop effective user state classification models even with mask cover.

Keywords: Deep Learning, Masks, Emotion Recognition

1 Introduction

This paper discusses the effect of mask wearing on the image processing-based user state detection methods used today. The goal is to ascertain the possible adverse effect of mask wearing to these systems and quantify the effect using arousal detection which is well studied in literature [1-5].

Many people worldwide work in conditions which require wearing of masks. Some professions like cleaning, maintenance work, mining and firefighting are done in debris or particle filled environments and therefore require specialized masks. Healthcare workers and laboratory workers wear medical grade masks to protect themselves from pathogens and other medical hazards. A lot of these professions are conducted under situations which require intense focus due to safety reasons. And the workers have to be in an optimal mental state in order to perform their tasks correctly and safely. With the increase of particulate matter in the air in large cities, as well as the advent of coronavirus pandemic, wearing of masks have become necessary for health reasons. With the normalization of mask wearing comes a host of challenges for image processing algorithms. In the emotion recognition domain, many of the state-of-the-art algorithms rely on facial image data [1,2] to do develop emotion classifiers. Features extracted from facial images [6-11] are then fed into a machine learning algorithm to develop a classification model and then use that model to classify new images. Facial occlusion

is particularly challenging in these approaches due to the sparse nature [14] of the feature matrices obtained. Even though there are suggested methods to deal with this issue such as filling in the gaps of the occluded images [12] or using the causal relations between facial regions [13], the need for hand crafted features has remained a challenge for adaptation of these methods.
Recent advances in deep learning have provided another opportunity to tackle the problem of facial occlusion by using automated feature extraction using CNN techniques [16, 17]. Such approaches have been obtaining state of the art results in the emotion recognition area [18]. This paper therefore leverages this technique in applying it to occluded facial images, specifically images with the lower part of the face occluded as in the case of mask wearing. The affect in the wild video dataset [2] is used as the basis of this work, and the images extracted from the video dataset were augmented to cover the lower part of face. The method is further expanded in the methods section. The results and implications of this work are discussed in the following sections.

2 Methods

There are many facial emotion datasets that are publicly available [19]. The affect in the wild dataset [2] was chosen for this work since, (1) it has a diverse set of faces. (2) the videos were captured in naturalistic settings. (3) in addition to the videos and affect annotations, it also contains 60 facial key points for each frame in each of the videos. This enabled placing an artificial occlusion on the images to simulate facial mask cover.

2.1 The dataset

The "affect in the wild" dataset consists of 298 videos, with a total length of more than 30 h. The videos were collected using the Youtube video sharing website using the keyword "reaction". The database displays subjects reacting to a variety of stimuli. Subjects display different combinations of emotions. The videos contain subjects from different genders and ethnicities with high variations in head pose and lighting. The videos were annotated using an online annotation procedure in which annotators used joysticks to provide arousal labels continuously (for each frame in the videos) in [−1, +1]. All subjects present in each video have been annotated. The total number of subjects is 200, with 130 of them being male and 70 of them female [2].

Dataset augmentation All the individual faces were cropped from each frame of the videos for the classification. The bounding box coordinate data from "affect in the wild" dataset was used for this. The annotation for each frame was applied to the individual faces as well. These individual face crops will be referred to as "dataset" from now onwards. Low and high labels were chosen as classification labels for the annotation ranges of [-1,-5] and [+5,+1]. The range [-5,+5] was dropped from the dataset due to the ambiguity in assigning high or low labels.

Fig. 1. Cropped faces from Affwild [3] dataset with associated labels.

This dataset was used as the benchmark for classification of faces without masks. Another dataset was created based on these images to represent the faces covered by masks. To achieve this, the facial key point data from "affect in the wild" dataset was used.

Fig. 2. Facial key points overlayed on a sample face.

The facial key points in the "affect in the wild" dataset ranged from 0-60. For this work, the key points from 2 to 15 were formed into a polygon shape to represent the facemask.

Fig. 3. Cropped faces with mask cover represented by a polygon using the facial key points of the lower half of face.

2.2 Training/testing configurations

(Test 1 – Across subject classifier without masks) Separated 20 videos from the without-mask-dataset and used the captured faces from those videos as the test set.
(Test 2 – Across subject classifier with masks) Separated 20 videos from the dataset-with-masks and used the captured faces from those videos as the test set.
(Test 3- Within subject classifier without masks) Separated 1000 faces of each label at random from the without-mask-dataset as the test set.

(Test 4- Within subject classifier with masks) Separated 1000 faces of each label at random from the dataset-with-masks as the test set.

For each of the above test methods, the train set, and test set were balanced to get the finalized dataset which contained approximately 40000 training images (20000 in low label and 20000 in high label) and approximately 2000 testing images (1000 in low label and 1000 in high label)

2.3 Preprocessing

The dataset images were rescaled to 180X180 before feeding into the classifier. They were converted to grayscale in order to avoid any bias from ambient color or lighting. Finally, they were normalized to the range of 0-1 from the pixel range of 0-255.

2.4 Classification

A RESNET50 [21] Convolutional Neural Network (CNN) classifier was used to train on the dataset and test on the separated test set. RESNET was chosen instead of a vanilla CNN due to the ability of RESNET to support larger number of layers. The batch size used for training was 32. And the validation split was 0.2. The loss was calculated using sparse categorical cross entropy. And the ADAM optimizer [20] was used with learning rate of 0.001, beta1 of 0.9, beta2 of 0.999 and epsilon of 1e-07. Early stopping was used to prevent overfitting the CNN model. The results from the two-label classification are described in the results section.

3 Results

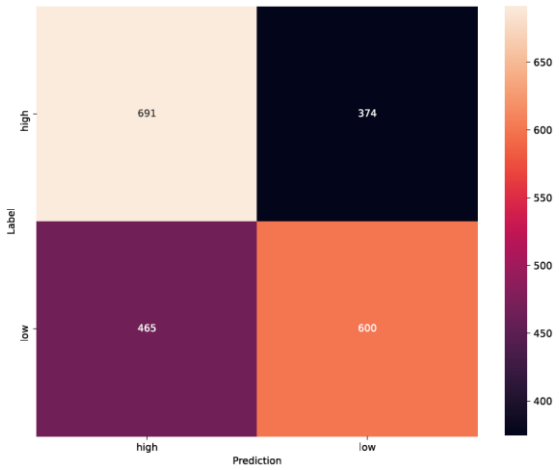

Fig. 4. Confusion matrix for Test 1 without mask overlay. Accuracy 61%

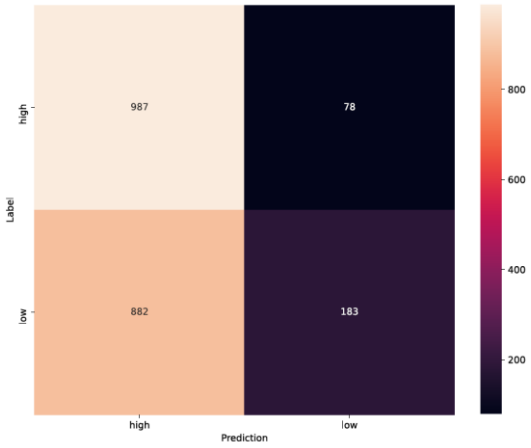

Fig. 5. Confusion matrix for Test 2 with mask overlay. Accuracy 55%

Table 1. Comparison of classification accuracy

Test	Mask cover	Accuracy
1	No	61%
2	Yes	55%
3	No	99%
4	Yes	97%

4 Discussion

The results (Table 1) show that test 1 and 2 obtained lower accuracy levels since the test set videos were not accessed at all by the classifier during the training phase. Test 3 and 4 achieved much better accuracy. This shows that arousal classification performs significantly better in within subject scenarios compared to across subject. Even though the results indicate that without mask dataset performs better, the accuracy could be higher with hyperparameter optimization. The main goal in this study was not to get the highest accuracy possible, but to offer a fair comparison between the dataset with and without mask cover.

In the test 1 and 2, using masks dropped the accuracy from 61% to 55%. This could be due to the lower half of the face contributing to the arousal detection, this will have to be confirmed by further studies looking at saliency maps from the model.
The accuracy level only dropped marginally (99% to 97%) between masked and non-masked case when it comes to within subject classification. Therefore, we can conclude that across subject classifiers will be more affected from mask cover than within subject classifiers. Arousal was the criteria used in this study. It will be interesting to see if the results translate to other forms of user state as well.
The results show that even with mask cover, it is possible to get acceptable level of accuracy for user state classification.

References

1. Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, *10*(1), 18-31.
2. D. Kollias, et. al.: "Deep Affect Prediction in-the-wild: Aff-Wild Database and Challenge, Deep Architectures, and Beyond". International Journal of Computer Vision (2019).
3. Machajdik, J., & Hanbury, A. (2010, October). Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM international conference on Multimedia* (pp. 83-92).
4. Nhan, B. R., & Chau, T. (2009). Classifying affective states using thermal infrared imaging of the human face. *IEEE Transactions on Biomedical Engineering*, *57*(4), 979-987.
5. Bandara, D., Velipasalar, S., Bratt, S., & Hirshfield, L. (2018). Building predictive models of emotion with functional near-infrared spectroscopy. *International Journal of Human-Computer Studies*, *110*, 75-85.
6. Buciu, I., Kotsia, I., & Pitas, I. (2003, November). Recognition of facial expressions in presence of partial occlusion. In *Proc. of the 9th Panhellenic Conference on Informatics (PCI'03)*.
7. Kotsia, I., Buciu, I., & Pitas, I. (2008). An analysis of facial expression recognition under partial facial image occlusion. *Image and Vision Computing*, *26*(7), 1052-1067.
8. Bourel, F., Chibelushi, C. C., & Low, A. A. (2001). Recognition of Facial Expressions in the Presence of Occlusion. In *BMVC* (pp. 1-10).
9. Zhang, L., Tjondronegoro, D., & Chandran, V. (2014). Random Gabor based templates for facial expression recognition in images with facial occlusion. *Neurocomputing*, *145*, 451-464.
10. Towner, H., & Slater, M. (2007, September). Reconstruction and recognition of occluded facial expressions using PCA. In *International Conference on Affective Computing and Intelligent Interaction* (pp. 36-47). Springer, Berlin, Heidelberg.
11. Huang, X., Zhao, G., Zheng, W., & Pietikäinen, M. (2012). Towards a dynamic expression recognition system under facial occlusion. *Pattern Recognition Letters*, *33*(16), 2181-2191.
12. Jiang, B., & Jia, K. B. (2011, September). Research of robust facial expression recognition under facial occlusion condition. In *International Conference on Active Media Technology* (pp. 92-100). Springer, Berlin, Heidelberg.

13. Miyakoshi, Y., & Kato, S. (2011, March). Facial emotion detection considering partial occlusion of face using Bayesian network. In *2011 IEEE Symposium on Computers & Informatics* (pp. 96-101). IEEE.
14. Cotter, S. F. (2010, March). Sparse representation for accurate classification of corrupted and occluded facial expressions. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 838-841). IEEE.
15. Hammal, Z., Arguin, M., & Gosselin, F. (2009). Comparing a novel model based on the transferable belief model with humans during the recognition of partially occluded facial expressions. *Journal of vision*, *9*(2), 22-22.
16. Roy, B., Nandy, S., Ghosh, D., Dutta, D., Biswas, P., & Das, T. (2020). MOXA: A Deep Learning Based Unmanned Approach For Real-Time Monitoring of People Wearing Medical Masks. *Transactions of the Indian National Academy of Engineering*, *5*(3), 509-518.
17. Loey, M., Manogaran, G., Taha, M. H. N., & Khalifa, N. E. M. (2021). Fighting against COVID-19: A novel deep learning model based on YOLO-v2 with ResNet-50 for medical face mask detection. *Sustainable Cities and Society*, *65*, 102600.
18. Bandara, D., Hirshfield, L., & Velipasalar, S. (2019). Classification of affect using deep learning on brain blood flow data. *Journal of Near Infrared Spectroscopy*, *27*(3), 206-219.
19. Buciu, I., Kotsia, I., & Pitas, I. (2005, March). Facial expression analysis under partial occlusion. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.* (Vol. 5, pp. v-453). IEEE.
20. Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
21. K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.