Computer Vision Project Proposal

Pranay Patil
University of Minnesota
patil122@umn.edu

Suraj Bandarupalli Ananya
University of Minnesota
banda057@umn.edu

1. Introduction

In recent years, online advertising spending has overtaken traditional media advertising spending. According to the Emarketer report 2019 [6], worldwide digital ad spending will rise by 17.6% to $333.25 billion. A. Guttmann [8], in the US, estimated the spending to amount to 283 billion U.S. dollars in 2018 and that it would further grow to 517 billion by the end of 2023. In the UK, China, Norway and Canada, digital advertisement is already the most dominant advertising medium. Tv advertisement, which was the leading advertising medium, has shown a steady decrease in spending. This change in spending has driven research for finding well-placed and personalized ads. According to the facial recognition study conducted by Annalect [1], the 'moodometer' study captured the facial expressions of 134 people as they watched five random Super Bowl 50 commercials of 2016. This data was used to rank the ads from most liked to least liked. The results were surprising as the study rated Mountain Dew's "Puppymonkeybaby" ad first while the same ad was ranked only 55th out of 63 ads on the USA Today Ad Meter list [20]. The ranking was incorrect as the moodometer study implemented the ranking based on positive and negative emotions but did not account for ad sentiment. A social message ad might get tear-jerking emotions but that does not necessarily mean that the ad was not liked. We propose a novel technique that not only looks at what emotion the user is feeling but also how engaged the user is and the ad's expected sentiment.

2. Related Work

2.1. Facial Emotion Recognition

Facial emotions are used to convey reactions and intentions of people to various stimuli. The objective is to derive the type of emotion which has been classified into 6 types: sadness,happiness, fear, anger, surprise and disgust from visual data. Later on Neutral was also added to this set. Initial algorithms first extracted features from the image and then used classifiers to detect emotion. Some examples of feature extraction are histogram of oriented gradients (HOG) [4], Gabor wavelets [13] and Haar features [22]. With the advancement in deep learning, Convolutional Neural Networks were used to in facial emotion recognition. It was a good fit as CNN can detect patterns from high and low level feature representations due to its multiple-layered architecture. Khorrami [12] achieved high levels of accuracy by using zero-bias CNN. To enhance generalizability for detecting facial emotion, Mollahosseiniet al. [15] trained CNN models across different well-known FER datasets. Liu in [14] used Boosted Deep Belief Network (BDBN) to achieve state-of-art accuracy. Han et al [9] proposed an incremental boosting CNN (IB-CNN) in order to improve the recognition of spontaneous facial expressions by boosting the discriminative neurons.

All the above methods have been successful in detecting facial emotions with very high accuracy. The decision of whether a user likes an ad lies beyond just facial emotion recognition. In this paper we propose a method to tackle the limitations of facial emotion recognition methods in online advertising.

2.2. Engagement Recognition

One of the initial research methods developed in detecting user engagement was by Kapoor [11] where different inputs like facial features, a pressure-sensitive mouse, a posture-sensing chair etc. was used to detect whether the student was frustrated. Grafsgarrdetal [7] worked on facial action units (AU) and linear regression methods to detect the relation between student engagement and AU. Whitehillet al. [23] classified four engagement levels: not engaged at all, nominally engaged, engaged in task, and very engaged using linear SVM, Gabor features and gentle boost. Boschet al. [3] detected engagement using AUs and Bayesian classifiers. Most of the above mentioned engagement recognition methods were performed in student engagement scenarios, where the length of engagement is longer than in the case of online advertising. It was also found by Whitehillet al. [23] that user engagement patterns are available in static images. Nezami [17] used CNN ,with weights set to VGG-B model trained for FER, to detect user engagement. The initial weights used lead to a performance boost in detecting unengaged scenarios. In our method, we intend to use the

same approach. In our method, we use a mathematical approach based on sclera and pupil regions in the eye to detect user's gaze.

3. Baseline method

As our baseline method we have chosen to use a facial emotion recognizer. We have selected to use a convolutional neural network (CNN) model to classify the facial image into 7 emotions, which are - angry, disgust, scared, happy, sad, surprised, neutral. Out of these 7 emotions, angry, sad and disgust are the emotions for our algorithm will switch the advertisement.

3.1. Dataset

For this naive approach, facial expression recognition (FER) data-set [10] from Kaggle was used. The data consists of 48x48 pixel grayscale images of faces. The faces have been automatically registered so that the face is more or less centered and occupies about the same amount of space in each image. The dataset consists of 35887 images which are labeled with 6 categories of emotions- (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral).

3.2. Approach

We trained a CNN with 4 blocks of convolution, batch normalization, max pooling and dropout layers, and at the end of which, it has 3 dense layers and a softmax layer, on the FER dataset [10].

The network has the total of 5,905,863 parameters out of which 5,902,151 are trainable. This model was trained with batch size of 32 and epoch value of 100. The approach for using this model in our advertisement problem is to first capture the frames of the user, who is in the frame. Then we use Haar feature-based cascade classifier [21] to detect the frontal face region. Which feature is then used to evaluate the person's emotion with our network, and if the maximum weighted emotion is one of the sad, disgust or angry our algorithm will switch the advertisement.

3.3. Results

The aforementioned network, gave the validation accuracy of 67.59% and the test accuracy of 64.83% for facial emotion detection. Due to lack of any dataset for human faces and their sentiment about an advertisement. We wouldn't measure the model's accuracy to accurately skip the advertisement. Confusion matrix for this network looks something like 1.

4. Proposed method

Inferring the viewer's sentiment towards an advertisement is not trivial. As proposed in the baseline method, we can't just rely on the viewer's facial emotions alone, but we should also take into consideration other factors such as viewer's engagement and advertisement's intended motive or sentiment. In this section we will briefly describe all the components of this method, and how can they be integrated to give better results than the baseline method3

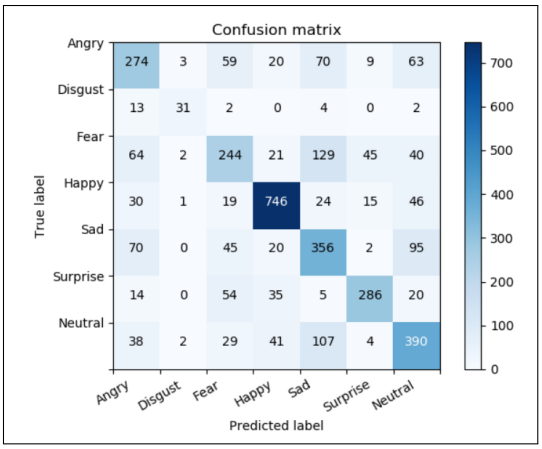

Figure 1. Confusion matrix

4.1. Facial emotion recognition

The most important component is detecting the facial emotions of the viewer. For this task we are using a convolutional neural network described in the paper [2], trained on the FER dataset [10]. The input to this network will be a facial landmark image, which can be obtained by the Haar feature-based cascade classifier [21], and the output will be a label vector containing the probabilities of each of the 7 emotions. The used model is based on Xception [5] architecture. It replaces vanilla convolution layers with depthwise separable convolution layers, and it also improves accuracy by adding residual modules and global average pooling in the network as shown in 2. This network outperformed the baseline approach 3 in terms of accuracy and performance. The proposed network has a validation accuracy of 70.69% and is 1.5x times faster.

4.2. Viewer engagement detection

Viewer's engagement plays crucial role in the experience of our approach. It is necessary to filter out viewer's emotions which are not results of the advertisement. The viewer might be talking to some other person or viewer's initial sentiment when he/she comes in the frame might be independent from her sentiment about the advertisement. Hence we are considering these external emotions as noise, and filtering those frames. The goal of this part in the proposed method is to detect whether the viewer is looking at the ad or not. While most of the above are good at detecting where the user is looking on the screen and decid-

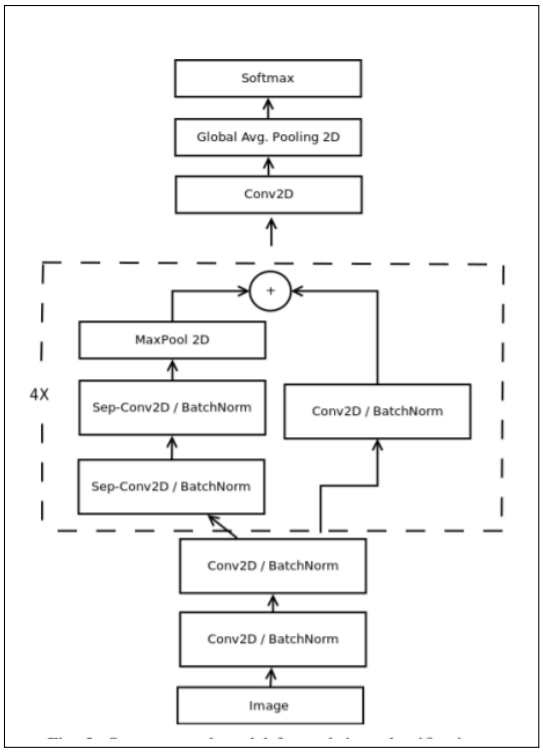

Figure 2. Proposed model for emotion recognition

ing user-engagement level using his/her emotions and viewpoint. While our goal is simpler, there is no good dataset to train a CNN model for viewer looking at the ad and not looking at the ad. Hence we decided to follow a mathematical approach. First we used a facial landmarks as template 3 to detect facial and eye regions. We use dlib library's face detector algorithm which is made using the classic Histogram of Oriented Gradients (HOG) feature combined with a linear classifier, an image pyramid and sliding window detection scheme. Once left eye region is detected, the eye region is split into 2 halves using center top and center bottom landmarks. Thresholding is done to the image to differentiate sclera(white part) from the pupil. The left-eye-ratio is taken between sclera on the left-side and right-side. Same process is followed for the right eye as well. An average of the two eye ratios is used to decide whether the user is looking at the ad or not. The calibration of threshold of gaze-ratio for looking outside versus inside was done manually and might vary for different screen sizes.

4.3. Advertisement sentiment detection

To accurately measure and quantify the viewer's reaction towards an advertisement, it is crucial to take advertisement's inherit emotion into consideration. For example, in the case of an movie poster for an horror movie should impart a scary reaction on the viewer, or an advertisement

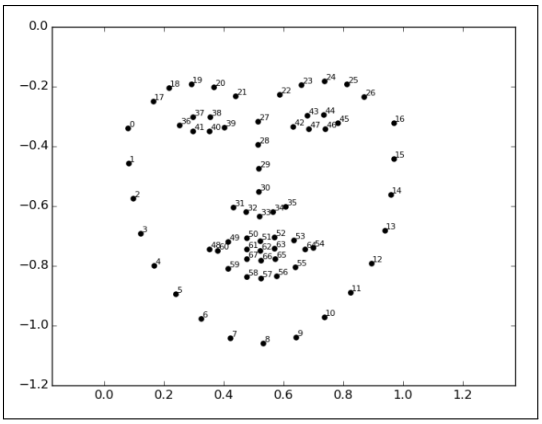

Figure 3. Facial landmarks template

about an no-profit environmental organization having image of the burning amazon should convey a sad emotion on the viewer. These negative emotions do not mean that the viewer has a dislike for this content, but he/she is in agreement with the advertisement. The baseline approach will penalize the model in such similar cases as it will consider sad or angry to be negative emotions. To compute emotion labels for an image we have trained a CNN with VGG-16 architecture on the deep emotion dataset [18]. This module is only needed when the expected advertisement sentiment is unknown and can be skipped for the cases where we can manually encode the sentiment vector for an advertisement. We achieved accuracy of 36%. This is due to the lack of accurately labeled dataset. The existing datasets only consist few thousands of images, which are not labeled very precisely.

4.4. Combining pieces

Components described in 4.1, 4.2 and 4.3 were built and combined and a demo application was built which was empowered by this pipeline. The goal of the project was to build an advertisement display system which will skip or won't skip the advertisement based on the viewer's reaction towards it. The 4.2 component could be used to filter out noisy frames i.e. frames in which viewer is not engaged with the content. After this filtering phase, the frame could be passed to the 4.1 component which will give us the viewer's emotion vector. This emotion vector and an advertisement emotion vector obtained from the component 4.3, can then be used to compute similarity between themselves. If this similarity score is below a certain threshold, then only we will change the displayed advertisement. Cosine similarity is used for calculating similarity, and the threshold value of 0.4 was selected by manual calibrations. 4 demonstrates the execution flow of the demo application which was built using the proposed approach. 5 displays the snippet of the application. There are three windows - adver-

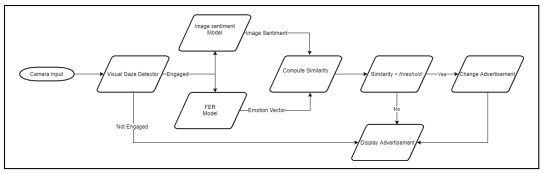

Figure 4. Execution flow

Figure 5. Demo application

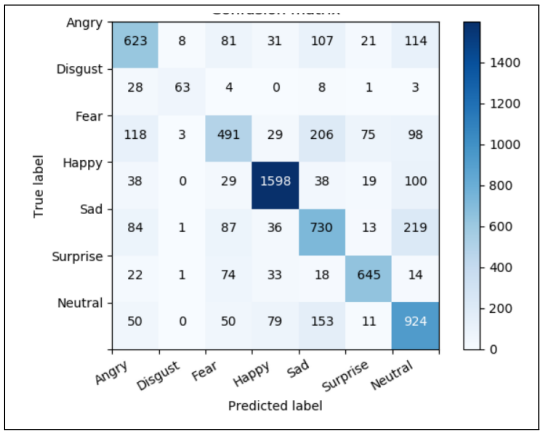

Figure 6. Confusion matrix

tisement, camera frame and facial emotion vector. The red label on the advertisement image represents the expected sentiment of the image, which matches the facial emotions and hence in this case the advertisement image will not be changed.

5. Result

Result of the proposed FER model and gaze detection algorithm can be seen in 6 and 7 respectively. For the analysis of our proposed method, we took two approaches. A Qualitative analysis which showcases the scenarios where our model outperforms the baseline approach and a Quantitative comparison where we take a sample of 100 combinations and test our method against the baseline.

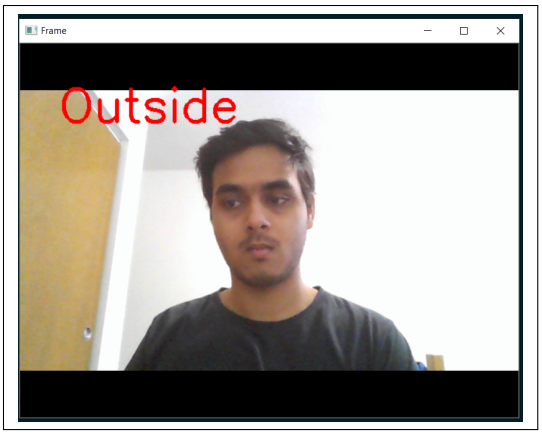

Figure 7. Gaze detection module

		Image sentiment		
		Positive	Negative	Total
Engagement	Positive	50	30	80
	Negative	10	10	20
	Total	60	40	100

Table 1. Sample distribution

5.1. Quantitative comparison

To make a quantitative comparison between the proposed and the baseline approach, we take a sample of 100 combinations of advertisement images and viewers engagement. Distribution of these samples can be seen in the table Sample distribution 5.1. We will compare three approaches on this data sample. Here we are considering the images have pre-labeled sentiment vectors for simple and clearer comparison.

1. Baseline approach

2. Proposed approach (only facial emotion recognition)

3. Proposed approach (engagement detection and facial emotion recognition)

The accuracy of correctly switching the advertisement depends on the accuracy of the FER network and number of images considered. For the first approach, the FER accuracy is 64.83% and it will consider all of the cases i.e. 100, even those where the viewer is not engaged. Other problem with this approach is it will always skip the ads for negative emotions, which is wrong. Hence the accuracy component for negative images is 0, and the total accuracy will depend on 50 positive and engaged images. For the second approach, the FER accuracy is 70.69% and it will consider all of the cases i.e. 100, even those where the viewer is not engaged. For the third approach, the FER accuracy is 70.69%, but it will only consider the cases where the viewer is engaged

Approach		Accuracy
1	$\frac{50 * 64.83}{100}$	32.42%
2	$\frac{80 * 70.69}{100}$	56.55%
3	$\frac{80 * 70.69}{80}$	70.69%

Table 2. Quantitative comparison

with the advertisement i.e. 80. The accuracies are computed in the table Quantitative comparison 5.1. It can be seen that, our approach outperforms the baseline method by 38.27

5.2. Qualitative result

Case 1: The gaze estimation model scans each frame of the input live feed for user engagement. We tested the framework for a few cases where the viewer showcases emotion towards something other than the ad. Our method detects that the user is not engaged and keeps displaying the ad irrespective of the emotion. The baseline method was successful if the user showed a positive emotion but changed the ad if the user was angry,sad or disgusted at something outside the ad. Both our method and baseline do not track emotions over a period of time and hence both the methods failed when the user showed impulsive emotions.
Case 2: We also tested for various combinations of ads and user emotions. As expected our method does not change the ad even though it is a negative emotion as long as the similarity between user sentiment vector and the ad sentiment vector is above the threshold. While this does out-perform the baseline, the accuracy of detecting emotion confidently plays a big role. In low illumination settings, the confidence in any one of the emotions is not high and hence the similarity is lower.

6. Conclusion

In this project we propose a framework to show more relevant and likable ads to user by using facial emotions, gaze and image sentiment. We showcase how our method outperforms a CNN facial emotion recognition model both qualitatively and quantitatively. For future work we would like to look at the following :

- Improving gaze detection and facial emotion recognition models for multi-face and low illumination scenarios
- Utilize RNN to get a more precise readings of facial emotions. This will be especially useful to detect and clear impulsive emotions that do not represent what the viewer is actually feeling

References

[1] annalect. Superbowl 2016, moodmeter, 2016.

[2] Octavio Arriaga, Matias Valdenegro, and Paul Plöger. Real-time convolutional neural networks for emotion and gender classification. 10 2017.

[3] Nigel Bosch, Sidney D'Mello, Ryan Baker, Jaclyn Ocumpaugh, Valerie Shute, Matthew Ventura, Lubin Wang, and Weinan Zhao. Automatic detection of learning-centered affective states in the wild. *International Conference on Intelligent User Interfaces, Proceedings IUI*, 2015:379–388, 03 2015.

[4] Junkai Chen, Zenghai Chen, Zheru Chi, and Hong Fu. Facial expression recognition based on facial components detection and hog features. 2014.

[5] Francois Chollet. Xception: Deep learning with depthwise separable convolutions. pages 1800–1807, 07 2017.

[6] Jasmine Enberg. Digital ad spending 2019, 2019.

[7] Joseph Grafsgaard, Joseph Wiggins, Alexandria Vail, Kristy Boyer, Eric Wiebe, and James Lester. The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring. pages 42–49, 11 2014.

[8] A. Guttmann. Digital advertising spending worldwide 2018-2023, 2019.

[9] Shizhong Han, Zibo Meng, Ahmed Shehab Khan, and Yan Tong. Incremental boosting convolutional neural network for facial action unit recognition. 07 2017.

[10] Kaggle. Challenges in representation learning: Facial expression recognition challenge, 2013.

[11] Ashish Kapoor, Winslow Burleson, and Rosalind Picard. Automatic prediction of frustration. *International Journal of Human-Computer Studies*, 65:724–736, 08 2007.

[12] Pooya Khorrami, Tom Paine, and Thomas Huang. Do deep neural networks learn facial action units when doing expression recognition? pages 19–27, 12 2015.

[13] Gwen Littlewort, Mark Frank, Claudia Lainscsek, Ian Fasel, and Javier Movellan. Recognizing facial expression: Machine learning and application to spontaneous behavior. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2:568–573, 06 2005.

[14] Ping Liu, Shizhong Han, Zibo Meng, and Yan Tong. Facial expression recognition via a boosted deep belief network. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1805–1812, 09 2014.

[15] Ali Mollahosseini, David Chan, and Mohammad Mahoor. Going deeper in facial expression recognition using deep neural networks. pages 1–10, 03 2016.

[16] Omid Mohamad Nezami, Mark Dras, Len Hamey, Deborah Richards, Stephen Wan, and Cécile Paris. Automatic recognition of student engagement using deep learning and facial expression. 2018.

[17] Omid Mohamad Nezami, Len Hamey, Deborah Richards, and Mark Dras. Deep learning for domain adaption: Engagement recognition. *CoRR*, abs/1808.02324, 2018.

[18] Hailin Jin Jianchao Yang Quanzeng You, Jiebo Luo. Building a large scale dataset for image emotion recognition: The fine print and the benchmark, 2016.

[19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[20] USA today. Usa today admeter results 2016, 2016.

[21] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. volume 1, pages I–511, 02 2001.

[22] Jacob Whitehill and Christian Omlin. Haar features for facs au recognition. pages 97–101, 01 2006.

[23] J. Whitehill, Z. Serpell, Y. Lin, A. Foster, and J. R. Movellan. The faces of engagement: Automatic recognition of student engagementfrom facial expressions. *IEEE Transactions on Affective Computing*, 5(1):86–98, Jan 2014.