Abstract—This paper presents a mobile solution for real
time facial emotion recognition running on an iPhone X. The
proposed system uses a neural network (NN) for classifying six
basic emotions. The facial expressions are extracted from 83
landmarks that are taken from specific vertices on a facial
mesh generated through Apple’s ARkit face geometry SDK.
The NN model achieves an accuracy of 80.42% solely on the
aforementioned 83 landmarks that are sourced from BU-3DFE
database. A mobile iOS application is created to demonstrate
the NN model processing feature inputs directly from the smart
phone. The real time mobile emotion recognition accurately
detects four of the basic emotions: Happy, Surprise, Anger,
Disgust (Figure 1).

I. INTRODUCTION

With front facing cameras and sensors becoming ubiq-
uitous within mobile devices, access to user facial data
has revealed an abundance of data that was previously not
easily available. Direct access to users faces presents the
opportunity for facial emotions to be extracted and analyzed
as users view content on their mobile devices. This data can
be used to gain various insight into how users feel while
interacting with products, such as, pain points within user
interfaces and, most importantly, user emotional satisfaction
within individual experiences.

Therefore, we propose a facial expression recognition
(FER) system that classifies the six universal facial emo-
tions, proposed by Paul Ekman [1], using Apple’s iPhone
X smartphone. Using a depth sensor, which is integrated
alongside the front facing camera of the iPhone, the facial
mesh produced is able to accurately track a user’s face. Facial
data, represented as 83 landmark vertices, are utilized as
input features into a mobile NN which subsequently outputs a
predicted user emotional state. As such, mobile device users’
emotional states are able to be unobtrusively gathered simply
through FER.

II. BACKGROUND

Facial expressions are difficult to classify due to the
variations between each individual person. However, studies
performed by psychologist Paul Ekman, which are cited
extensively within the field of affective computing, have
found that humans express six universal facial expressions
regardless of culture, age or gender, including, happiness,
sadness, anger, fear, surprise and disgust [1]. Many early
attempts at classifying these expressions developed classifiers
that utilized exaggerated facial models, however, these are
not an accurate depiction of the expressions humans typically
display. Kobayashi & Hara attempted to solve this problem
by developing a dual neural network (NN) system in which
one NN was trained on strong facial expressions, while the
latter was trained on weak facial expressions [2]. The dual
NNs were able to label each expression as weak or strong;
thus developing a discrete FER classifier that identified the
six universal emotions in a binary manner (i.e. weak or
strong). Moreover, the NNs were trained on facial models
within a laboratory setting, in which lighting and head
positioning could be perfectly controlled. This, however, is
not translatable to real world applications.

To develop a more robust system, S. Chickerur and K.
Joshi developed a lighting-impartial solution in which 3D
images were used to detect users portraits and profiles in
order to classify their expressions [3]. More specifically, a
Microsoft Kinect sensor was used to capture red, green and
blue (RGB) images and facial depth data for live classifica-
tion. This system solved the illumination problem by using
the depth component to guide the creation of facial meshes.
The facial meshes allowed important feature vertices to be
accurately compared to their training set. Moreover, the depth
data allowed the solution to classify facial expressions when
only the profile of the users face was exposed. Unfortunately,
the Microsoft Kinect sensor is cumbersome and uses too
much energy to effectively be used as a mobile solution.

III. PROPOSED SOLUTION

The solution is composed of three main components,
including, (1) a facial expression recognition system via the
TABLE I: Neural Network Architecture Comparison

<table>
<thead>
<tr>
<th>Model #</th>
<th>NN Structure</th>
<th>Hidden Layer Activation Function</th>
<th>Cost Function</th>
<th>Epochs</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>249-127-127-6</td>
<td>Leaky ReLu</td>
<td>Cross Entropy</td>
<td>10,000</td>
<td>77.92</td>
</tr>
<tr>
<td>2</td>
<td>249-127-127-6</td>
<td>ReLu</td>
<td>L2</td>
<td>10,000</td>
<td>80.42</td>
</tr>
<tr>
<td>3</td>
<td>249-249-249-6</td>
<td>Leaky ReLu</td>
<td>Cross Entropy</td>
<td>10,000</td>
<td>78.75</td>
</tr>
<tr>
<td>4</td>
<td>249-249-249-6</td>
<td>ReLu</td>
<td>L2</td>
<td>10,000</td>
<td>77.50</td>
</tr>
</tbody>
</table>

iPhones X, (2) emotion recognition through a Tensorflow neural network (NN), and (3) a mobile platform in which to integrate facial recognition and the NN model together (Figure 1). The result is an iOS application that interprets emotions through detected facial expressions.

A. Facial Expression Recognition

Apples most recently released smartphone, the iPhone X, was utilized to detect users facial expressions. The iPhone X, which contains a proprietary depth sensor integrated within the front-facing camera, allows for accurate facial detection. This feature, along with Apples iOS SDK, provides a framework with which to efficiently and consistently detect faces and extract their features. The iOS SDK, and specifically the ARkit package, is equipped with a suite of utilities related to Augmented Reality (AR) and Facial Geometry (FG) recognition. Therefore, using the aforementioned packages, an iOS application was developed which utilized depth sensor data to generate a single facial mesh over each users face using a consistent number of vertices. The generated facial mesh consistently identifies the same 1,220 vertices for each new face that is detected. However, only 83 landmark vertices were utilized to detect emotions, as described in more detail below. The chosen landmarks correlate to the features identified within the BU-3DFE database, which contains 100 subjects ranging in age and ethnicity [4].

![Fig. 2: Applying BU-3DFE landmarks to ARkit facial mesh](image)

B. Emotion Recognition

A neural network (NN) was developed within TensorFlow (TF) and trained to correlate the detected facial expressions with specific user emotions. The BU-3DFE database was utilized as ground truth training data for the NN. Each subject within the database was pictured with varying degrees of each of the six Ekman emotions (e.g. level 1 happy, level 2 happy, etc.), resulting in 2400 total 3D facial models [4]. Each image within the database was labeled with the 83 feature landmarks, and each landmark consisted of the X, Y and Z components. Therefore, 249 input features (83 x 3 axis) were utilized to train the NN. Several NN architectures were tested to compare which combination of parameters provided the most accurate results, as shown in Table I.

TABLE II: One-Hot Vector Classes

<table>
<thead>
<tr>
<th>Emotion</th>
<th>One-Hot Vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happy</td>
<td>[1, 0, 0, 0, 0, 0]</td>
</tr>
<tr>
<td>Anger</td>
<td>[0, 1, 0, 0, 0, 0]</td>
</tr>
<tr>
<td>Sad</td>
<td>[0, 0, 1, 0, 0, 0]</td>
</tr>
<tr>
<td>Surprise</td>
<td>[0, 0, 0, 1, 0, 0]</td>
</tr>
<tr>
<td>Fear</td>
<td>[0, 0, 0, 0, 1, 0]</td>
</tr>
<tr>
<td>Disgust</td>
<td>[0, 0, 0, 0, 0, 1]</td>
</tr>
</tbody>
</table>

Ultimately, Model 2, from Table I, resulted in the highest accuracy (80.42%). Model 2 is composed of a 249-feature input layer, two hidden layers comprised of 127 nodes each, and a six-node output layer, representing the six classes of emotions (Figure 3). Each hidden layer utilizes the rectified linear unit (ReLU) activation function. This function is known for increasing NN training speeds due to its monotonic nature and because it does not suffer from the vanishing gradient problems. The final output layer uses the softmax activation function in order to distribute the NN predictions so that they add up to 1; allowing for a categorical representation of the results using a one-hot vector representation (Table II). The L2 (i.e. least squared error) cost function was used to determine the prediction error after each epoch. Finally, the Adam optimizer was utilized to optimize weights and improve model accuracy.

C. Mobile Platform

Past literature involving FER research most commonly utilized the Microsoft Kinect to obtain facial depth data [5], [6]. However, the Kinect is a stationary platform and does not effectively work in low light settings, thus it can not be used to create a mobile emotion recognizer. The iPhone X alleviate these problems through its inherent mobility and its use of a flood illuminator, allowing the depth sensor to function in the dark. These features made the iPhone X an ideal candidate for a mobile platform.

Additionally, TensorFlow was utilized for model creation due to its available conversion tools. Specifically, the conversion tools within TF allow NN models to be converted to alternate platforms, such as the iPhone X. Finally, Apple
maintains a machine learning SDK, coreML, which allows NN models to efficiently run on mobile devices directly. Therefore, the model was able to be efficiently trained on a powerful computer before converting to a coreML format and integrated into an iOS application. In summary, the final iOS application utilizes 83 specific FG landmarks as inputs to a trained NN model within a mobile platform, thus allowing users to determine current emotions from facial expression data.

IV. RESULTS

The developed iOS application demonstrates the feasibility of emotion detection solely through the recognition of 83 vertex landmarks gathered from a facial mesh. Emotions are predicted using a CoreML NN model, which was converted from TF, and utilizes the normalized facial feature landmarks as inputs to the model.

A. Facial Expression Recognition

Each of the 83 landmark vertices designated on the facial mesh were manually correlated with the 83 landmarks within the BU-3DFE database. Due to this, certain facial landmarks were difficult to assign properly, specifically the eyebrows and the mouth. The BU-3DFE database has landmarks placed along the shape of each facial feature (e.g. eyebrows, lips, etc.), such that landmarks are customized to every individual model. This non-generalization of landmark placement lends it difficult to manually designate vertices on the facial mesh. Improving landmark placement so that they better align with the BU-3DFE landmarks, may allow for better emotion classification but at the risk of reducing generalization.

B. Emotion Recognition

Currently, the application accurately detects emotional states of happiness, surprise, anger, and disgust, as shown in Figure 4. The system must have a confidence level greater than 90% to output the predicted emotion, otherwise, a neutral classification is output. Fear and sadness are much more difficult for the system to detect due to many of their landmark positions overlapping with anger and disgust, causing the models confidence to remain below the 90% threshold necessary to make a final classification.

V. CONCLUSION

This paper presents an efficient and consistent mobile approach to facial emotion recognition through the iPhone X platform. The suggested system employs accurate facial landmark detection using ARkits constant face mesh generation. Since no auxiliary processing is necessary to place the landmark on the mesh this phase of facial recognition processing remains fast and efficient. The landmarks are utilized as inputs to a CoreML NN model which correlates the positioning of the landmarks with one of the six Ekman facial emotions. The final result is a mobile iOS application that can recognize user emotions from a facial mesh generated with Apples depth sensor. The proposed solution performs efficiently and detects user emotions to within 80% accuracy.

Further improving the accuracy of the NN Model would allow for this system to be utilized within a variety of mobile applications, such as content platforms and mobile games. For example, the system would allow for analysis of users emotional state when interacting with an application and improve the user experience accordingly. Specifically, user emotional data may be used to determine pain points within a user interface, enabling developers to update their applications to provide a more satisfying product. Furthermore, this data may be used to help identify which content (e.g. videos, images, etc.) causes users to positively emote, thus providing a large advantage to advertisers. Finally, user emotions may be utilized within gaming platforms, allowing games to adjust in real-time to better challenge and/or reward the player. Overall, the applications of this mobile emotion recognition engine are vast and may be used to improve how affective computing is implemented in day-to-day applications.
VI. FUTURE WORK

Increasing overall accuracy of the NN Model would allow this system to be effectively used in enterprise software. This may be achieved through additional hyperparameter configuration testing while examining the effect of each change on the overall system. Furthermore, the NN model could be converted to a predictive system, in which emotions are determined prior to the visual cue occurring. This approach would require the use of the BU-4DFE dataset (or a similar database) in conjunction with an alternate NN architecture (e.g. Recurrent or LSTM). The NN would train on videos, as opposed to images, of landmarked faces to determine temporal landmarks that occur prior to emotions visually occurring within the face.

REFERENCES

[4] Lijun Yin; Xiaozhou Wei; Yi Sun; Jun Wang; Matthew J. Rosato, A 3D Facial Expression Database For Facial Behavior Research, 7th International Conference on Automatic Face and Gesture Recognition, 10-12 April 2006 P:211 - 216