Low-compute facial expression recognition using fiducial feature-sets

A.V. Divekar and D.C. Gharpure
Dept. of Electronic Science, Savitribai Phule Pune University, Pune, India
E-mail: ankur.divekar@gmail.com, dcgunipune@gmail.com

Abstract. Facial Expression Recognition is an exciting area of affective computing. As mobile and embedded devices become increasingly ubiquitous, exploration of low-compute approaches for facial expression recognition is essential. Facial landmark points are fiducial features that are used to localize and represent salient regions of the face, such as eyes, nose and lips. Any facial expression can be expressed as an activation of facial muscles in specific parts of the face, thereby affecting the locations of the facial landmark points in those parts. This relationship can be captured concretely by deriving appropriate feature-sets from these points. In this paper, an approach for deriving three types of feature-sets from a set of facial landmark points detected on a face is discussed. Such feature-sets are derived for three standard facial expression recognition datasets with labelled expression classes. The derived feature-sets are used as inputs for training computationally light machine-learning classifiers, yielding encouraging classification accuracy. These results are presented and discussed with both quantitative and qualitative observations.

1. Introduction
Facial expressions are an important modality of non-verbal communication during human interaction and can convey a wide array of emotions and intentions. For several years now, automated facial expression recognition (FER) has attracted significant research interest and activity in the field of computer vision. FER has been applied to numerous applications such as advanced driver assistance systems, gaming and virtual reality. Numerous promising approaches that utilize deep neural networks (DNNs) for FER have been reported. Training DNNs is a highly energy-intensive process with high computation requirements, and thereby less than optimal for large-scale deployments. On the other hand, an increasing number of mobile and embedded devices now run algorithms suitable for their relatively lower, but significant computational abilities. This makes the exploration of low-compute approaches essential, especially for applications like FER, which may be deployed on low-end devices at extremely large scales. Numerous commercial phone apps utilize facial landmark points (FLPs) for digital manipulation of facial images. Naturally, FLPs are a logical starting point for FER applications too.

In this paper, we present an approach that utilizes facial landmark points to derive fiducial feature-sets. We report the FER classification performance of multiple algorithms with low-compute requirements. The details of datasets used are given in section 3.1. In order to train
and test the facial expression classifier, the images in the dataset must be translated into a feature vector that can be applied as an input to the classifier. In sections 3.2 and 3.3, the process of deriving the fiducial feature-sets for each facial image has been described. Section 4.1 describes the process of training different classifiers using the derived feature-sets. Finally, section 4.2 provides details of the FER performance evaluation and observations thereof.

2. Previous work
A number of approaches towards FER have been demonstrated, using various hand-crafted features such as Local Binary Patterns, Singular Value Decomposition, Bag of Words (BoW), Histogram of Gradients (HoG), and Scale Invariant Feature Transform (SIFT). These methods have shown encouraging results on several standard databases.[1]

Numerous methods for estimating and localizing FLPs have been developed[2]. Various FER approaches that utilize facial geometric features have also been previously reported[5]. FLPs have been utilized as the intermediate step for numerous face-related tasks such as face alignment and facial image synthesis, apart from expression recognition.

3. Methodology
Ekman et. al. identified the six universal facial expressions[8] as anger, disgust, fear, happiness, sadness and surprise. Combined with a neutral expression, this creates seven fundamental, distinct classes of facial expression. A machine-learning system can be built to classify a given facial image into one of these pre-defined classes as shown in Figure 1

An appropriately labeled dataset with numerous instances of each expression class is required to train a machine-learning algorithm. End-to-end training approaches like convolutional neural networks generally utilize raw image data as input. However a classifier with significantly lower compute requirements can be built by performing feature extraction on the facial image. For a given facial image, the first step is to estimate the location of several facial landmark points which represent salient regions of the face such as eyebrows, eyes, nose and lips. A set of fiducial features is then derived by quantifying the positions of these facial landmark points relative to a common reference point. A feature vector representing various measurements of the face is thus extracted. The set of such extracted features from a labeled dataset is used to train a classifier. This trained classifier may then be utilized to make a prediction of expression class for an image that has not been seen by the classifier before.

3.1. Datasets
Curated datasets are one of the driving factors behind rapid progress in fields using machine learning and deep learning. Three standard FER datasets were utilized for the experiments. The details of these databases are as follows:

(i) ADFES: The Amsterdam Dynamic Facial Expression Set[9] has been created at the Amsterdam Interdisciplinary Centre for Emotion. The dataset consists of videos as well as still images of 12 male and 10 female subjects with 10 facial expressions. The still images were obtained by isolating a single video frame at the apex point of the model’s expression.
(ii) TFEID: The Taiwanese Facial Expression Image Database[10] has been jointly by the Brain Mapping Laboratory at National Yang-Ming University and the Integrated Brain Research Unit at Taipei Veterans General Hospital in Taiwan. Only a subset of the entire dataset has been made available so far for research purposes. This subset consists of images of 20 male and 20 female subjects with eight posed facial expressions, captured in frontal view.

(iii) WSEFEP: The Warsaw Set of Emotional Facial Expression Pictures[11] dataset contains images from 14 male and 16 female models with 7 facial expressions. The images were carefully selected to fit criteria of basic emotions and then evaluated by independent judges. In order to maintain consistency within the training data, only front-facing images were utilized, and images captured at other angles were ignored. Further, only images for the 7 expressions common to all three datasets were selected in order to minimize categorical imbalance. Data augmentation was done by executing a horizontal flip on every selected image, and adding the flipped image to the dataset, thereby doubling the number of samples in each expression class, as shown in Table 1.

Table 1: Expression-label counts by dataset

| Expression | ADFES | TFEID | WSEFEP | Total |
|------------|-------|-------|--------|-------|
| Neutral    | 44    | 78    | 60     | 182   |
| Anger      | 44    | 68    | 60     | 172   |
| Disgust    | 44    | 80    | 60     | 184   |
| Fear       | 44    | 80    | 60     | 184   |
| Joy        | 44    | 80    | 60     | 184   |
| Sadness    | 44    | 78    | 60     | 182   |
| Surprise   | 42    | 72    | 60     | 174   |
| Total      | 306   | 536   | 420    | 1262  |

While the three datasets were also utilized individually for the experiments, the samples of each expression class were merged across the three datasets to form a composite dataset, which was used to examine the feasibility of training a classifier capable of generalizing across racially diverse faces.

3.2. Feature Extraction

In order to ensure uniformity of extracted features, the facial images must be preprocessed before feature extraction. For each image in the database, facial detection was performed. The detected face was then aligned and resized to predefined dimensions. 68 Facial Landmark Points(FLPs) were localized on the aligned and resized facial image, as illustrated in Figure 2b. A set of such FLPs was generated for all the images in the database. Given this set of FLPs, feature extraction was performed through the following steps:

Step 1: Let $F$ be a set of $n$ chosen FLPs. Then for image $i$ in the dataset, the individual FLPs can be denoted as $f_j^i = (x_j, y_j)$ where $x_j$ is the x-coordinate, $y_j$ is the y-coordinate and
\( j \) is the index number of the FLP, from 1 to \( n \). The centroid \( C_{x,y}^i \) of all the FLPs is defined as

\[
C_{x,y}^i = \left( \frac{\sum_{j=1}^{n} x_i}{n}, \frac{\sum_{j=1}^{n} y_i}{n} \right)
\]

Step 2: Let the length of the line joining an FLP \( f_j \) to \( C_{x,y} \) be denoted as \( l_j \). Then,

\[
l_j = \sqrt{(x_j - C_{x}^i)^2 + (y_j - C_{y}^i)^2}
\]

Let \( L \) be the set of the respective lengths of all the lines joining each individual FLP to the centroid. Then,

\[
L = [l_1, l_2, l_3, \ldots, l_n]
\]

Let \( l_{\text{max}} \) denote the maxima of the set of lengths. Then the set of normalized lengths is represented as

\[
L_{\text{norm}}^i = \left[ \frac{l_1}{l_{\text{max}}}, \frac{l_2}{l_{\text{max}}}, \frac{l_3}{l_{\text{max}}}, \ldots, \frac{l_n}{l_{\text{max}}} \right]
\]

Step 3: Next, consider a coordinate system with origin at the centroid calculated earlier. Then the angle \( a_j \) created between x-axis and the line joining FLP \( f_i \) to the centroid is calculated as

\[
a_j = \arctan \left( \frac{y_j - C_{y}^i}{x_j - C_{x}^i} \right)
\]

Let \( A \) be the set of all angles thus calculated for the \( n \) FLPs. Then,

\[
A = [a_1, a_2, a_3, \ldots, a_n]
\]

Let \( a_{\text{max}} \) and \( a_{\text{min}} \) be the maxima and minima of \( A \) respectively. Then the set of angles normalized between 0 and 1 is obtained as:

\[
A_{\text{norm}}^i = \left[ \frac{a_1 - a_{\text{min}}}{a_{\text{max}} - a_{\text{min}}}, \frac{a_2 - a_{\text{min}}}{a_{\text{max}} - a_{\text{min}}}, \frac{a_3 - a_{\text{min}}}{a_{\text{max}} - a_{\text{min}}}, \ldots, \frac{a_n - a_{\text{min}}}{a_{\text{max}} - a_{\text{min}}} \right]
\]

Step 4: The individual sets of length and angle features extracted for image \( i \) are \( L_{\text{norm}}^i \) and \( A_{\text{norm}}^i \) respectively.

Step 5: The composite feature vector \( AL^i \) for image \( i \) is created by concatenating the sets of normalized lengths and angles. Thus,

\[
AL^i = [L_{\text{norm}}^i, A_{\text{norm}}^i]
\]

Step 6: Let the set of length feature vectors for all the \( n \) images in the dataset be \( L_n \). Similarly \( A_n \) and \( AL_n \) are the sets of angle and concatenated feature vectors respectively.
3.3. Feature-set Design

It is observed that face detection systems often generate a crop of the facial image that excludes the outer periphery of the facial area, such as the top of forehead and the chin. To check the feasibility of using these cropped images for FER, a subset of the 68 detected FLPs was also considered. The FLPs for the chin, the jaw-line, and the side of the face were excluded, to simulate a tightly cropped facial image. The remainder subset consisted of 46 FLPs for left eyebrow, left eye, right eyebrow, right eye, nose and mouth.

Considering the two sets of FLPs (68 and 48 FLPs) and the two types of features (length and angle) extracted, six feature-sets were created as follows:

(i) $A_{68}$: Angle features $A_{\text{norm}}$ for the 68-FLP set.
(ii) $L_{68}$: Length features $L_{\text{norm}}$ for the 68-FLP set (as shown in Fig. 2c).
(iii) $AL_{68}$ Concatenated angle and length features for the 68-FLP set.
(iv) $A_{46}$: Angle features $A_{\text{norm}}$ for the 46-FLP subset.
(v) $L_{46}$: Length features $L_{\text{norm}}$ for the 46-FLP subset (as shown in Fig. 2d).
(vi) $AL_{46}$ Concatenated angle and length features for the 46-FLP subset.

4. Experiments

The 6 feature-sets represent different types of fiducial measurements of the face. In order to examine the classification accuracy of each feature-set, they were tested on the 3 standard databases as well as the composite database described in 3.1.

4.1. Training and Testing

For sufficiently large datasets, it is often observed that multiple machine-learning classifiers are able to learn the internal patterns of the data, and give consistent and nearly equivalent results. However, for relatively small datasets such as the one utilized here, the appropriate choice of classifier algorithm can create a significant difference in the prediction. Hence, three classifiers that are known to be particularly efficient for small and medium-sized datasets were specifically selected for the experiments:

(i) Random Forest classifier[12]
(ii) Support Vector Machine(SVM)[13]
(iii) XGBoost Classifier[14]
In order to get average accuracy figures for each classifier, k-fold cross validation was utilized, with \( k=5 \). The input feature-set for each experiment run was accordingly split into training(80%) and test(20%) sets. Each of the classifiers was fit on the training set, and then the predictions were generated for the test set and used to calculate accuracy for that experiment. As per \( k=5 \), the experiment was repeated 4 more times with different train-test splits of the feature-set, and the accuracies were averaged. This process was repeated for all six feature-sets, and the results obtained are tabulated in the next section.

4.2. Results
Table 2 lists the classification accuracies obtained for the three classifiers on all six feature-sets derived from the ADFES dataset. It is observed that the accuracy for for \( L_{68} \) is better than \( A_{68} \) for two of the three classifiers. Additionally the accuracy for \( AL_{68} \) is actually less than \( L_{68} \). It may be concluded that the angle features do not contribute anything significant to the classification and perhaps only add noise to the feature-set with 68 FLPs. On the other hand, it can be observed that the average accuracy of \( L_{46} \) is less than that of \( A_{46} \), and the angle features seem to have the greater contribution to the accuracy of \( AL_{46} \).

| Classifier | \( A_{68} \) | \( L_{68} \) | \( AL_{68} \) | \( A_{46} \) | \( L_{46} \) | \( AL_{46} \) |
|------------|--------|--------|--------|--------|--------|--------|
| RandomForest | 0.85   | 0.81   | 0.81   | 0.85   | 0.81   | 0.85   |
| SVM        | 0.81   | 0.85   | 0.81   | 0.85   | 0.78   | 0.81   |
| XGBoost    | 0.85   | 0.85   | 0.85   | 0.81   | 0.78   | 0.89   |

Table 3 lists the classification accuracies obtained for the three classifiers on all six feature-sets derived from the TFEID dataset. Similar to the ADFES accuracy figures, it can be observed that length features have higher accuracies than angle features for the feature-set with 68 FLPs. In the case of the feature-set with 46 FLPs, the angle features have low accuracy by themselves, but exhibit a significant contribution to the accuracy of \( AL_{46} \).

| Classifier | \( A_{68} \) | \( L_{68} \) | \( AL_{68} \) | \( A_{46} \) | \( L_{46} \) | \( AL_{46} \) |
|------------|--------|--------|--------|--------|--------|--------|
| RandomForest | 0.76   | 0.76   | 0.8    | 0.76   | 0.72   | 0.76   |
| SVM        | 0.80   | 0.83   | 0.87   | 0.76   | 0.76   | 0.83   |
| XGBoost    | 0.76   | 0.83   | 0.83   | 0.74   | 0.76   | 0.85   |

Table 4 lists the classification accuracies obtained for the three classifiers on all six feature-sets derived from the WSEFEP dataset. A similar trend is observed for this dataset too. Angular
features have lower accuracy than length features, but clearly contribute to the accuracy of the composite feature vector, particularly in the feature-set with 46 FLPs.

Table 4: Classification Accuracy for WSEFEP

| Classifier  | feature-set | A68 | L68 | AL68 | A46 | L46 | AL46 |
|-------------|-------------|-----|-----|------|-----|-----|------|
| RandomForest |             | 0.75| 0.81| 0.86 | 0.81| 0.83| 0.89 |
| SVM         |             | 0.78| 0.86| 0.89 | 0.72| 0.92| 0.86 |
| XGBoost     |             | 0.83| 0.86| 0.86 | 0.75| 0.75| 0.83 |

Table 5 lists the classification accuracies obtained for the three classifiers on all six feature-sets derived from the composite dataset created by merging the samples of each expression class in the three datasets above. Some very interesting observations may be made. The length features clearly outperform the angular features with all three classifiers in both the 68-FLP feature-sets as well as 46-FLP feature-sets. However, the more striking improvement in classification accuracy is seen in the AL68 as well as AL46. This clearly indicates that training a classifier on multiple racially diverse databases actually helps the classifier to generalize better, and reduces overfitting to faces with a particular set of visual traits. The high accuracy of AL46 also indicates that FER may be successfully implemented on images with a close-cropped facial area.

Table 5: Classification Accuracy for composite dataset

| Classifier  | feature-set | A68 | L68 | AL68 | A46 | L46 | AL46 |
|-------------|-------------|-----|-----|------|-----|-----|------|
| RandomForest |             | 0.77| 0.86| 0.91 | 0.78| 0.86| 0.94 |
| SVM         |             | 0.84| 0.93| 0.94 | 0.85| 0.91| 0.95 |
| XGBoost     |             | 0.80| 0.87| 0.92 | 0.81| 0.83| 0.91 |

5. Conclusion
As the use of systems executing face-related analysis becomes increasingly wide-spread, it is essential to explore methods that can be successfully deployed at scale without humungous power and computation budget requirements. Feature extraction based on fiducial points is a great candidate in this regard, as facial landmark point detection is well-tested and proven on fairly low end commercial-off-the-shelf hardware such as cell-phones and embedded boards like Raspberry Pi. In this paper, an analysis of FER using feature-sets generated through fiducial measurements based on facial landmark points has been presented. The classification accuracy has been evaluated on multiple datasets individually and collectively. It is observed that higher classification accuracy can be obtained by training a classifier on a diverse dataset. While the performance of the SVM classifier on the individual datasets is mixed, it scores the highest accuracy for the composite dataset. Notably, the fact that AL46 scores higher accuracy
than $AL_{68}$ seems to indicate that judicious exclusion of certain features can actually reduce the noise in the dataset, and help better classifier generalization. Further work on statistical feature selection in order to choose appropriate features is being studied. Additionally, methods of optimizing the classification accuracy as a function of specific expression classes are being explored, with respect to specific applications. While the current accuracy figures are lower than those obtained from deep-learning approaches, they are nevertheless encouraging from the perspective of large-scale deployment for consumer applications, considering their humble computational requirements.

References

[1] Roychowdhury S and Emmons M 2015 International Journal of Computer Science & Engineering Survey ISSN 09763252
[2] Johnston B and de Chazal P 2018 A review of image-based automatic facial landmark identification techniques
[3] Kotsia I and Pitas I 2007 IEEE Transactions on Image Processing 16 172–187 ISSN 10577149
[4] Ghimire D and Lee J 2013 Sensors (Switzerland) 13 7714–7734 ISSN 14248220
[5] Majumder A, Behera L and Subramanian V K 2014 Pattern Recognition vol 47 pp 1282–1293 ISSN 00313203
[6] Kowalski M, Narumiec J and Trzciński T 2017 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops vol 2017-July pp 2034–2043 ISBN 9781538607336 ISSN 21607516 (Preprint 1706.01789)
[7] Yang X, Li Y, Qi H and Lyu S 2019 IH and MMSec 2019 - Proceedings of the ACM Workshop on Information Hiding and Multimedia Security pp 113–118 ISBN 9781450368216 (Preprint 1904.00167)
[8] Ekman P and Friesen W V 1971 Journal of Personality and Social Psychology ISSN 00223514
[9] van der Schalk J, Hawk S T, Fischer A H and Doosje B 2011 Emotion 11 907–920 ISSN 1931-1516 URL http://doi.apa.org/getdoi.cfm?doi=10.1037/a0023853
[10] Chen L and Chen Y 2007 Taiwanese facial expression image database
[11] Olszanowski M, Pochwatko G, Kuklinski K, Scibor-Rylski M, Lewinski P and Ohme R K 2014 Frontiers in Psychology ISSN 16641078
[12] Prokhorenkova L, Gusev G, Vorobev A, Dorogush A V and Gulin A 2018 Advances in Neural Information Processing Systems ISSN 10495258 (Preprint 1706.09516)
[13] Cortes C and Vapnik V 1995 Machine Learning ISSN 15730565
[14] Chen T and Guerstrin C 2016 Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining ISBN 9781450342322 (Preprint 1603.02754)