2019 IEEE Symposium Series on Computational Intelligence (SSCI)
December 6-9 2019, Xiamen, China

Mutual-Information-based Feature Selection for Facial Emotion Recognition on Light-Weight Devices

Yingjun Dong
Department of System Science and Industrial Engineering
State University of New York at Binghamton
Binghamton, New York, 13902, USA
Email: ydong25@binghamton.edu

Hiroki Sayama
Department of System Science and Industrial Engineering
State University of New York at Binghamton
Binghamton, New York, 13902, USA;
Waseda Innovation Lab Waseda University
Shinjuku, Tokyo 169-8050, Japan
Email: sayama@binghamton.edu

***Abstract*—Light-weight devices have become ubiquitous in our daily life, such as smartphones, smart monitors, and other smart devices in our home. As light-weight devices are becoming popular, the demand for sophisticated human-computer interaction (HCI) applications for light-weight devices is also increasing. One particularly promising HCI application for light-weight devices is facial expression recognition (FER), since it may open up possibilities of various medical, psychological or psychiatric monitoring. However, its high computational demand has prevented widespread adoption of FER on light-weight devices. To address this issue, here we aim at decreasing computational overhead of FER by reducing the number of facial landmarks. We calculated mutual information of facial landmarks' movements and detected their clusters using hierarchical agglomerative clustering (HAC). We also applied a genetic algorithm (GA)-inspired landmark selection method to filter out low-utility features from each facial landmark cluster. The selected features were provided to a support vector machine (SVM) classifier to classify facial expressions, and its performance was compared among several different algorithm settings. Results showed that our proposed method achieved classification accuracy similar to the classifier that used the original full-featured dataset, with improved performance robustness and computational time reduced by 63.5%.**

***Keywords*—facial expression recognition; light-weight devices; feature selection; mutual information; hierarchical agglomerative clustering; support vector machine.**

I. Introduction

With the development of 5G technology, light-weight devices, especially mobile phones, are becoming increasingly common and useful in our daily life. As such light-weight devices decrease in price and become more widely available, they occupy an increasingly critical position in human life, in both personal/entertainment and professional/business scenes [1]. Several advanced applications of human-computer interaction (HCI) have been developed for light-weight devices. For example, some smartphones now use facial recognition to unlock the system.

One promising HCI technology that has significant potential benefit is facial expression recognition (FER). Over the past decades, FER has been utilized in several successful applications, including analysis of human emotional behaviors and monitoring of patients' emotional status in hospitals. Typical FER methods are computationally demanding using 25–130 landmarks [2]–[6], and therefore, many of the earlier FER studies utilized high-end stand-alone computational environments. Meanwhile, implementation of FER on mobile devices is also actively studied because of its prospect to realize greater flexibility and convenience. Earlier studies in this direction implemented FER on high-end mobile devices with substantial computational power [7], [8]. There is still a gap in this body of literature regarding how to implement FER on more computationally limited light-weight devices that are more widely available on the market, without losing recognition performance. Reducing computational overhead will also help reduce power consumption, leading to more continuous, more robust FER on those devices.

In this study, we aim to reduce the number of facial landmarks required in FER by detecting informational correlations among them and carefully selecting the most useful features from the correlated feature clusters. We used a dataset obtained from the Manual Annotation on AR Face Database [6], [9], which contains 2D coordinates of 130 manually annotated facial landmarks for 112 subjects' four different facial expressions. The movements (displacements) of facial landmarks' coordinates across different facial expressions represent a coordinated unique pattern of facial muscle behaviors. Such landmark movements contain a lot of information about facial expressions and thus were used to classify facial expressions. We measured mutual information (MI) between pairs of facial landmarks with regard to their movements. Using the results of MI calculation, we constructed a MI distance matrix. We then applied hierarchical agglomerative clustering (HAC) to the matrix to classify the landmarks into clusters of similar movement patterns. We selected one representative feature from each cluster using several different methods, and then constructed a support vector machine (SVM)-based facial expression classifier. We evaluated the performance of the developed classifiers through comparison with the classifier that used the original full-featured dataset.

978-1-7281-2485-8/19/$31.00 ©2019 IEEE

The rest of the paper is structured as follows. In Section II, we will review the relevant literature. We will discuss detailed methods in Section III. The design and results of experiments will be described in Sections IV and V, respectively. Finally, Section VI concludes the paper with future research directions.

II. Related Work

A. Feature Selection with Mutual Information

Feature selection is a critical problem in pattern recognition. It is used to remove redundant and irrelevant features and thereby improve the performance of pattern recognition. Mutual information (MI), i.e., information-theoretic nonlinear correlation between two random variables [10], has been heavily utilized in feature selection studies to detect correlations between features and outcome variables. A classic is the work by Battiti [11] that introduced MI for feature selection (MIFS), which selects features based on their MI with class variables. This method can reduce the dimensionality of input data, and it is now considered an important procedure for classification [12].

MIFS was followed by a large number of studies that improved its performance. Peng et al. [13] used the maximal statistical dependency criterion based on MI to select good features. As there were difficulties in calculating maximal dependency directly, they developed the minimal-redundancy-maximal-relevance criterion (mRMR), which led to smaller classification errors. Based on mRMR, Zhang et al. [14] proposed the mCRE method that included mRMR, clustering, and recursive feature elimination. The mCRE method was shown to choose fewer features with higher classification accuracy. Estevez et al. [15] proposed normalized MI feature selection (NMIFS) that used normalized MI as a measure of redundancy. Yin et al. [16] proposed improved normalized MI feature selection (INMIFS) by introducing a new quality estimation function. The INMIFS method shows good results both in accuracy and redundancy reduction. More recent developments in MI-based feature selection include Lee et al.'s work [17] on multi-label feature selection and Gao et al.'s work [18] on dynamic changes of selected features (DCSF) using conditional MI between selected features and classes.

These prior studies commonly used MI to measure correlations between features and variables to be explained or predicted. In our study, in contrast, we will use MI between features themselves to identify informational clusters that we can exploit for reduction of the number of features, which is different from the earlier works reviewed above.

In addition to feature selection on information theory, there are other methods for measuring similarity. Yu et al. [19] created a feature selection model named Fast Correlation Based Filter (FCBF). They measured F-correlations between pairwise features, and used symmetrical uncertainty as the goodness measure. Their approach was successful in reducing computational overhead and removing redundant and irrelevant features. Zhang et al. [20] conducted a hybrid feature selection algorithm, in which they applied one-class F-score, improved F-score and genetic algorithm to do feature selection. Then, they applied four classification methods, k-nearest neighbors (k-NN), random forest, Gaussian naïve Bayes and SVM, to evaluate the selected features. Their work demonstrated improved performance, but time efficiency was not considered.

Genetic algorithm (GA) is also a widely used feature selection method. Vafaie et al. [21] compared the results of image texture recognition using sequential backward selection (SBS) with those of GA. They found the features selected by GA worked better than those selected by SBS. Oh et al. [22] developed a hybrid GA for feature selection, which made some changes in local search operations based on the typical GA. They improved offsprings with local search operations applied before the replacement step. In this way, they controlled the size of offsprings while improving the overall performance.

In this paper, we present a new feature selection application that was not explored in the literature reviewed above. We propose a hybrid feature selection method for selecting a small number of facial landmarks for facial expression recognition tasks. Our approach is based on the information theoretic analysis of the characteristics of the data.

B. Facial Expression Recognition

Over the past decades, facial expression recognition (FER) has become a major research area with significant achievements, including FER based on Gabor Wavelets [23]–[25] and FER by local binary pattern (LBP) [26]. Facial landmark localization is an essential part of FER as well [3]–[6], in which most studies detected and used at least 25 (and often a lot more) facial landmarks. Real-time FER [2] is another hot research topic, especially on mobile platforms. Choi et al. [27] developed locally random incremental classifier (LRIC) using local random projection (LRP) to extract facial features in real time efficiently. Suk et al. [7] developed a smartphone FER application using SVM and the Active Shape Model (ASM). Suchitra et al. [28] proposed a real-time FER method for mobile devices using the Haar cascade and ASM.

In these prior studies, the number of facial landmarks used for FER remained fairly large, ranging from 26 to 77. Naturally, the number of facial landmarks is directly linked to the computational overhead of FER, and our aim is to reduce it significantly.

III. Methods

The overall methodology is summarized in Fig. 1. The whole work was done using Python 3. We will describe details of the method in the following part.

A. Facial Landmarks' Movement

The dataset we used was obtained from the Manual Annotation on AR Face Database [9], which includes 112 subjects' 130 facial landmark coordinates in four different facial expressions (neutral, smile, anger and scream) recorded on two different days. We set the neutral expressions' coordinates as the baseline $(x_{i,j}^{\mathrm{neu},\delta}, y_{i,j}^{\mathrm{neu},\delta})$, where $\delta \in \{1, 2\}$ is the day of recording and i and j are indices for subjects and facial

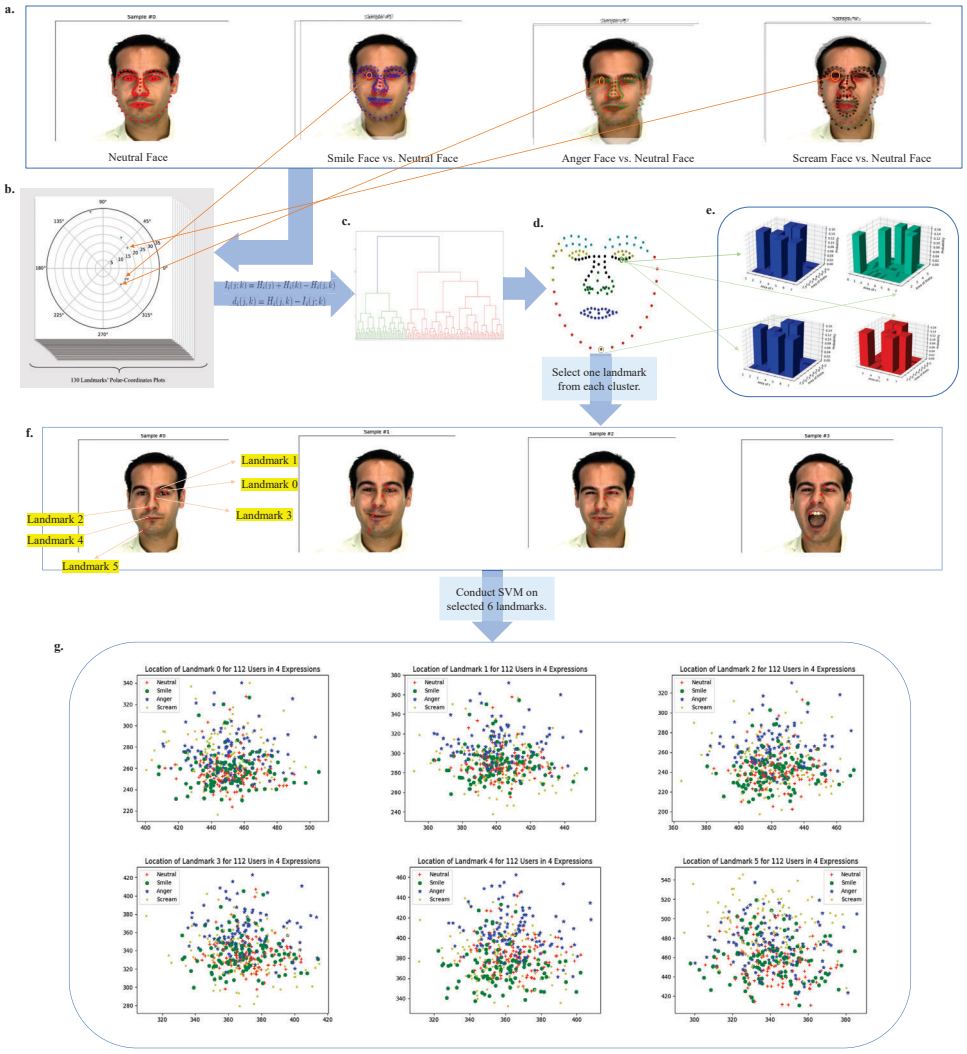

Fig. 1. A schematic illustration of our proposed method. **a.** We set the locations of 130 facial landmarks on the neutral expression as the baseline and measured the displacement of the landmarks in the other three expressions. **b.** We binned the movements of each landmark in polar coordinates to obtain its movement distribution. We obtained 130 such distributions for each subject's face. **c.** We calculated mutual information and mutual information distance between every pair of movement distributions and conducted the hierarchical clustering of the landmarks using their mutual information distances. **d.** The result of the clustering. **e.** Landmarks in the same cluster have similar movement distributions while landmarks in different clusters have distinct movement distributions. **f.** We selected one landmark from each cluster. **g.** Using the selected 6 landmarks, we classified different expressions with SVM.

landmarks, respectively. We then measured the landmarks' displacements from the baseline for each of the other expressions and captured them in a polar coordinate system, as follows:

$$\Delta x_{i,j}^{\epsilon,\delta} = x_{i,j}^{\epsilon,\delta} - x_{i,j}^{\text{neu},\delta} \tag{1}$$

$$\Delta y_{i,j}^{\epsilon,\delta} = y_{i,j}^{\epsilon,\delta} - y_{i,j}^{\text{neu},\delta} \tag{2}$$

$$r_{i,j}^{\epsilon,\delta} = \sqrt{(\Delta x_{i,j}^{\epsilon,\delta})^2 + (\Delta y_{i,j}^{\epsilon,\delta})^2} \tag{3}$$

$$\phi_{i,j}^{\epsilon,\delta} = \text{atan2}(\Delta y_{i,j}^{\epsilon,\delta}, \Delta x_{i,j}^{\epsilon,\delta}) \tag{4}$$

Here, $\epsilon \in \{\text{smi (smile)}, \text{ang (anger)}, \text{scr (scream)}\}$ is the facial expression. $(\Delta x_{i,j}^{\epsilon,\delta}, \Delta y_{i,j}^{\epsilon,\delta})$ represents the displacement of facial landmark j of subject i in expression ϵ on day δ in a Euclidean coordinate system, and $(r_{i,j}^{\epsilon,\delta}, \phi_{i,j}^{\epsilon,\delta})$ represents the same displacement in a polar coordinate system that has a better rotational symmetry. These polar coordinate data were binned into multiple discrete bins by unit distance (5 in this study) for r and unit angle (45° in this study) for ϕ so that $r \in \{0\text{-}5, 5\text{-}10, \ldots, \}$ and $\phi \in \{0\text{-}45°, 45\text{-}90°, \ldots\}$. Using these discretized data, we constructed $p_{i,j}(r, \phi)$, a discrete probability distribution of movements of subject i's facial landmark j (Fig. 2), and also $p_{i,j,k}(r_j, \phi_j, r_k, \phi_k)$, a joint probability distribution of simultaneous movements of subject i's two landmarks j and k.

B. Mutual Information and Distance

Using $p_{i,j}$ and $p_{i,j,k}$ obtained above, we calculated Shannon's information entropies, joint entropies and mutual infor-

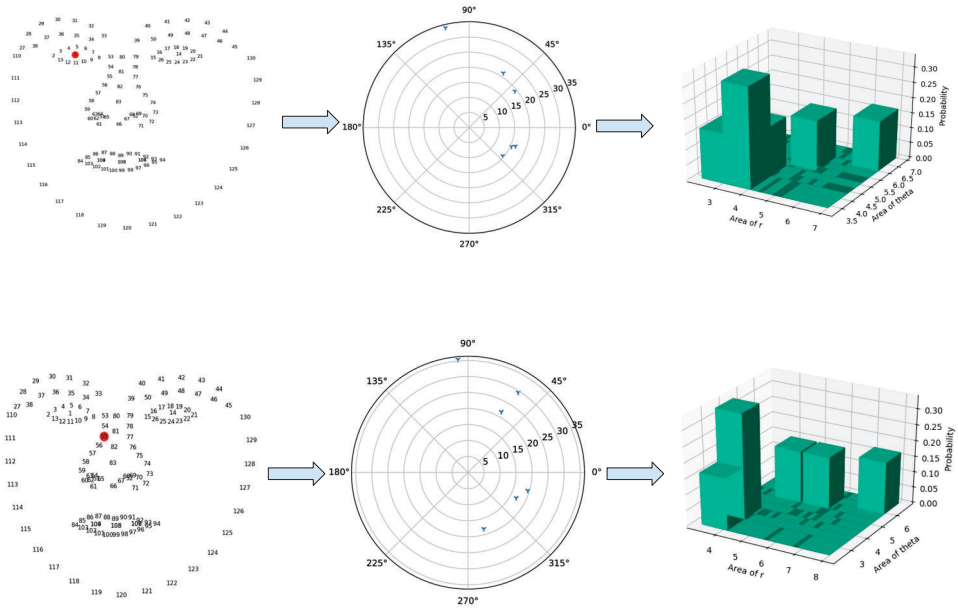

Fig. 2. Examples of probability distributions of facial landmark movements $p_{i,j}(r, \phi)$. Red dots in face figures on the left show which landmark was visualized. The polar scatter plots in the middle show distributions of the landmark's movements. The histograms on the right show the probability distributions in the (r, ϕ) space. Top: Right iris. Bottom: Right side of the nose root. Different facial landmarks show different, though potentially correlated, movement distributions.

mation of landmarks' movements as follows:

$$H_i(j) = -\sum_{r,\phi} p_{i,j}(r,\phi) \log p_{i,j}(r,\phi)$$

$$H_i(j,k) = -\sum_{\substack{r_j,\,\phi_j,\\ r_k,\,\phi_k}} p_{i,j,k}(r_j,\phi_j,r_k,\phi_k) \log p_{i,j,k}(r_j,\phi_j,r_k,\phi_k)$$

$$I_i(j;k) = H_i(j) + H_i(k) - H_i(j,k)$$

The MI values tell us which pairs of subject i's facial landmarks have higher nonlinear correlations in movements.

To conduct clustering on landmarks, we converted MI values into MI distance metrics, as follows:

$$\begin{aligned} d_i(j,k) &= H_i(j,k) - I_i(j;k) \\ &= H_i(j) + H_i(k) - 2I_i(j;k) \end{aligned}$$

These MI distance metrics were organized into a 130×130 distance matrix $D_i = (d_i(j,k))_{j,k}$ for subject i. In total, we obtained 112 such distance matrices ($n = 112$), which were then averaged for all subjects as follows:

$$\bar{D} = \frac{1}{n} \sum_{i=1}^{n} D_i$$

This average matrix $\bar{D}$ gives a symmetric table of average MI distances for every pair of landmarks obtained from all subjects, which is visualized in Fig. 3.

C. Hierarchical Agglomerative Clustering

We conducted hierarchical agglomerative clustering (HAC) on the average MI distance matrix $\bar{D}$ to detect clusters of

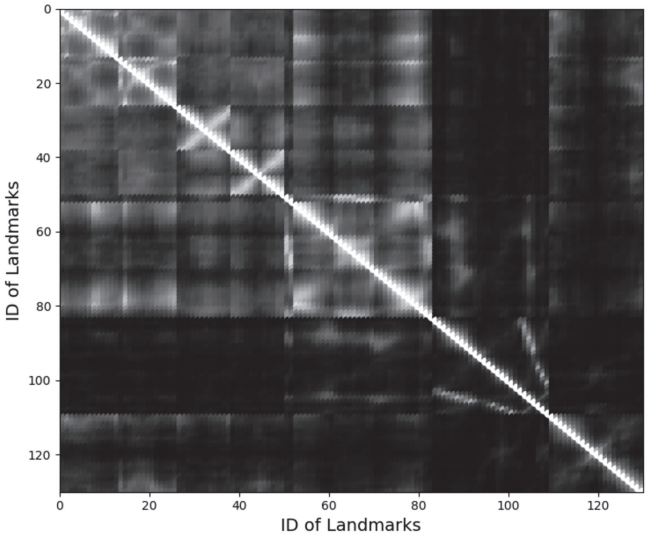

Fig. 3. Visualization of the average MI distance matrix $\bar{D}$. X-axis and y-axis represent the IDs of landmarks j and k, respectively. The shade represents the value of MI distance between a pair of landmarks (light = small, dark = large). The diagonal white line shows self-identity with distance 0. There are some cluster structures already visible in this figure.

landmarks' movements. We applied Ward's method [29] in HAC. The resulting dendrogram is shown in Fig. 4. We applied the Thorndike method [30] to decide the number of clusters, with the minimum required number of clusters set to 5. Fig. 5 shows the distances between clusters joined, in which the

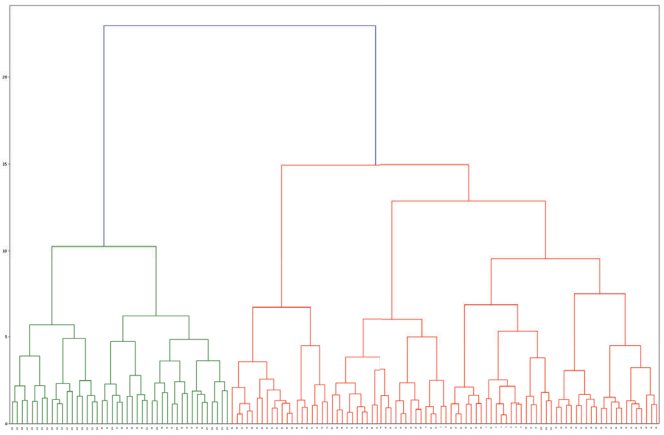

Fig. 4. HAC dendrogram. From bottom to top shows the agglomeration process by which individual landmarks were gradually combined into one cluster. X-axis represents the ID of landmarks, while y-axis represents the distance between the joined clusters.

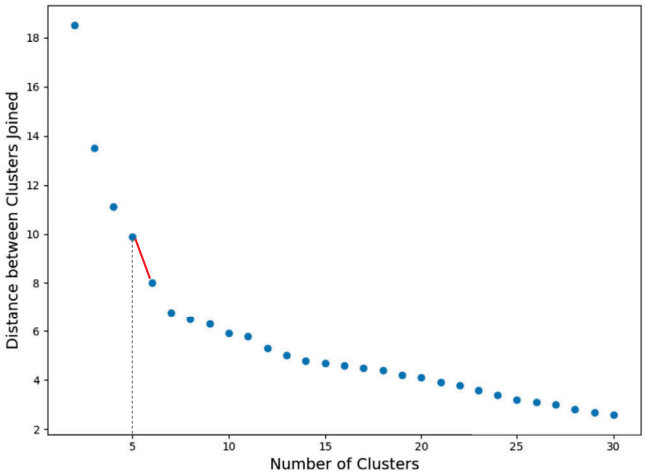

Fig. 5. Distances between joined clusters plotted over the number of clusters. The dotted line in this figure shows the minimum required number of clusters assumed in this study (5). The red line shows the biggest gap (between 5 and 6), so we chose 6 as the number of clusters.

biggest gap was observed between 5 and 6. Based on this result, we chose 6 as the number of clusters. The spatial distributions of these 6 clusters of facial landmarks are as shown in Fig. 6.

D. Support Vector Machine Classification

We used the support vector machine (SVM) [31] as a classifier of facial expressions. The clusters of facial landmarks obtained above were utilized to select representative landmarks in several different ways (details will be explained in the next section). The coordinates of selected facial landmarks were used as inputs to SVM, and the classification model was built to predict which facial expression the subject was showing. A radial basis function (RBF) was used as a kernel function in this SVM classification.

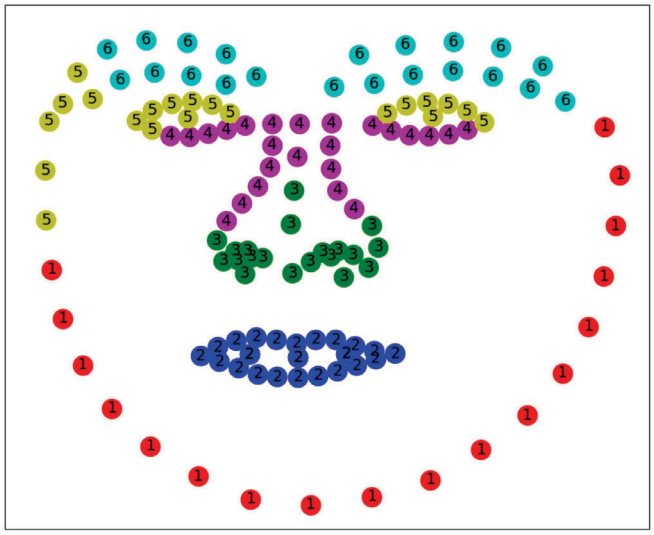

Fig. 6. Visualization of 6 facial landmark clusters detected in this study. Different colors/numbers represent different clusters.

E. Landmark Selection

In our experiments, we selected landmarks in two ways. One is a random selection (either from the whole set of 130 landmarks or from each of the six clusters), and the other is to filter landmarks using a computationally light selection process inspired by genetic algorithm. We call the latter selection method "FF" (for feature filtering) hereafter.

The FF landmark selection was implemented as follows:

1) Put the 130 landmarks into m bin(s). In this study, we used either $m = 1$ (i.e., all the 130 landmarks are in a single bin) or $m = 6$ (i.e., the landmarks are separated according to the result of clustering).
2) Choose $6/m$ landmark(s) randomly from each bin to create a set of 6 landmarks.
3) Conduct the SVM classification using the 6 landmarks selected above, and evaluate its accuracy.
4) Repeat 2 and 3 above t times and keep only the results whose accuracies exceed threshold θ. In this study, we used $t = 1000$ and initial threshold $\theta = 0.87$.
5) Refresh the contents of the bins by removing landmarks that did not appear in any of the results that met the accuracy threshold above.
6) Increase θ a little and repeat 2–5 above.

Through these steps, landmarks that are not useful for facial expression recognition will be quickly eliminated. We tested this landmark selection method with several variations of settings and found that conducting the filtering (steps 2–5 above) just for one iteration would produce results as good as those of multiple iterations, and therefore, all the results presented in this paper are based on a single iteration of filtering.

IV. Experiments

A. Feature Selection Methods

The following five feature selection methods were implemented and compared with each other in terms of classification performance and computational time.

1) 130 landmarks
   - This method uses all the 130 landmarks included in the data for classification with no feature selection. This method serves as the baseline.
2) R6
   - This method uses 6 landmarks that are randomly selected from the original 130 landmarks.
3) S6
   - This method uses 6 landmarks, each of which is randomly selected from one of the 6 clusters.
4) FF on R6
   - This method uses 6 landmarks that are selected from the original 130 landmarks after FF is applied.
5) FF on S6
   - This method uses 6 landmarks, each of which is selected from one of the 6 clusters after FF is applied.

B. SVM Classification and Computational Time

We used each of the above five methods to generate a dataset of the landmarks' coordinates that were then used for classification with SVM. Each dataset was split into a training set and a test set with the splitting rate of 20%. Specifically, we used 90 subjects' data as the training set and 22 subjects' data as the test set.

We ran the classification 1,000 times on each of the five datasets, and then we compared their accuracy results.

We used Raspberry Pi 3 as a representative of light-weight devices in our study. The Raspberry Pi 3 we used was a B+ model with a quad-core processor.

V. Results

The accuracy results of the classification are shown in Fig. 7. We can see that the accuracies of datasets of S6 and FF on S6 are better than those of R6 and FF on R6. This means that the landmarks selected from clusters outperformed the landmarks not selected from clusters. We also compared the results of S6 and FF on S6 against the result obtained using 130 landmarks. We found that the result of S6 worked slightly worse than that of 130 landmarks, but the result of FF on S6 worked as good as (even slightly better than) that of 130 landmarks, and moreover, the results of FF on S6 were more consistent with less variance than those of 130 landmarks. To show the comparison directly, we listed the statistical values of the accuracy results in Table I.

We conducted the Mann-Whitney U test on the classification results of S6 and R6. The p-value was 1.4127×10^{-19}, which means distributions of accuracy results in S6 and R6 are significantly different. We also conducted the same test on the classification results of FF on S6 and FF on R6. The p-value was 1.4796×10^{-32}. We concluded that the distributions of accuracy results in FF on S6 and FF on R6 are significantly different.

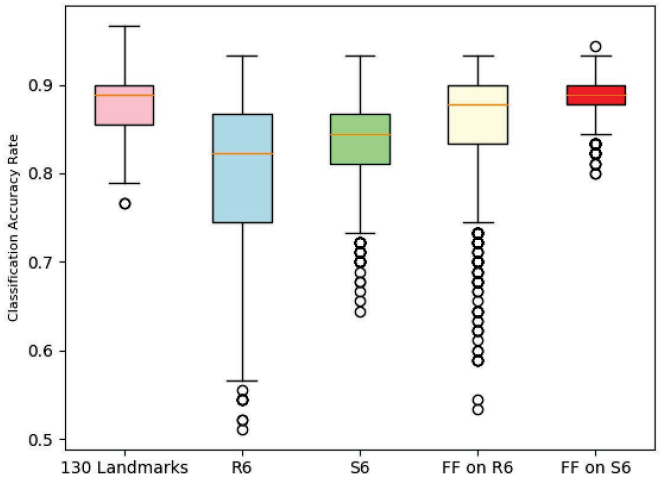

Fig. 7. Comparison of accuracy results of running SVM classification 1,000 times on different feature selection methods.

TABLE I
Statistics Table of Accuracy Results

Datasets	Mean	Standard Deviation	(Minimum, Maximum)
130 Landmarks	0.8821	0.0326	(0.7667, 0.9667)
R6	0.7998	0.0820	(0.5111, 0.9333)
S6	0.8366	0.0450	(0.6444, 0.9333)
FF on R6	0.8558	0.6258	(0.5333, 0.9333)
FF on S6	0.8876	0.0225	(0.8, 0.9444)

The computational time distributions on Raspberry Pi are shown in Fig. 8. The computational time of running classification experiments with dataset of selected 6 landmarks was reduced by 63.5% compared to the computation time with the dataset of 130 landmarks.

VI. Conclusion and Future Work

This study sheds new light on mitigating computational overhead on light-weight devices. We proposed a feature selection method with measuring the relationship on facial landmarks and clustering the facial landmarks to select 6 landmarks. We achieved good performance in classification and computational time reduction with selected landmarks. We concluded that FF on S6 shows more robust performance even compared to 130 landmarks, since FF on S6 showed less variance of accuracies. Using a small number of facial landmarks in FER is significant to reduce computational time. However, the scale of the dataset we used in our study is small, and we will find more datasets to do the work in the future.

In the future, we will conduct feature selection on datasets of 3D facial landmarks. We will find 6 or more landmarks which are the most representative to create a pre-trained facial landmarks detection model. We will apply the selected 6 facial

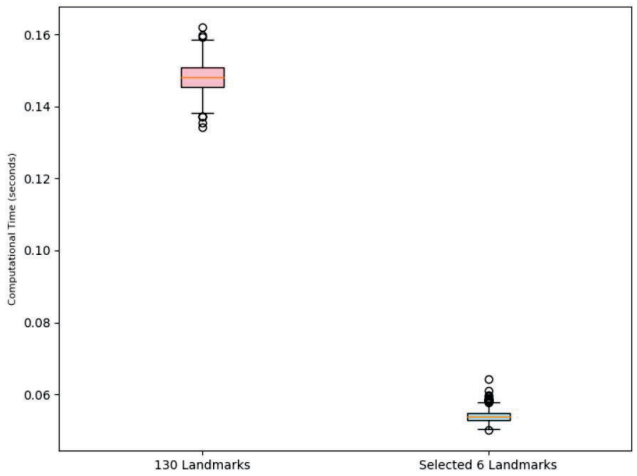

Fig. 8. Comparison of computation time for running 1,000 SVM classifications on Raspberry Pi.

landmarks for other tasks, such as tracking facial landmarks' movement in a video.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant #1734147.

REFERENCES

[1] H. Falaki, R. Mahajan, S. Kandula, D. Lymberopoulos, R. Govindan, and D. Estrin, "Diversity in smartphone usage," in *Proceedings of the 8th international conference on Mobile systems, applications, and services*. ACM, 2010, pp. 179–194.

[2] R. El Kaliouby and P. Robinson, "Real-time inference of complex mental states from facial expressions and head gestures," in *Real-time vision for human-computer interaction*. Springer, 2005, pp. 181–200.

[3] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 12, pp. 2930–2940, 2013.

[4] Y. Chang, C. Hu, and M. Turk, "Probabilistic expression analysis on manifolds," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, 2004.

[5] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image Vision Comput.*, vol. 28, no. 5, pp. 807–813, 2010. [Online]. Available: http://dx.doi.org/10.1016/j.imavis.2009.08.002

[6] A. M. Martinez and R. Benavente, "The ar face database," *24 CVC Technical Report*, 1998.

[7] M. Suk and B. Prabhakaran, "Real-time mobile facial expression recognition system-a case study," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 132–137.

[8] D. McDuff, A. Mahmoud, M. Mavadati, M. Amr, J. Turcot, and R. e. Kaliouby, "Affdex sdk: a cross-platform real-time multi-face expression recognition toolkit," in *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*. ACM, 2016, pp. 3723–3726.

[9] L. Ding and A. M. Martinez, "Features versus context: An approach for precise and detailed detection and delineation of faces and facial features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 2022–2038, 2010.

[10] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[11] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on neural networks*, vol. 5, no. 4, pp. 537–550, 1994.

[12] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, pp. 1157–1182, 2003.

[13] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[14] Q. Zhang, H. Wang, and S. W. Yoon, "A hierarchical feature selection model using clustering and recursive elimination methods," in *Proceedings of the 2017 Industrial and Systems Engineering Research Conference*, 2017.

[15] P. A. Estévez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," *IEEE Transactions on Neural Networks*, vol. 20, no. 2, pp. 189–201, 2009.

[16] L. Yin, M. Xingfei, Y. Mengxi, Z. Wei, and G. Wenqiang, "Improved feature selection based on normalized mutual information," in *2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)*. IEEE, 2015, pp. 518–522.

[17] J. Lee and D.-W. Kim, "Mutual information-based multi-label feature selection using interaction information," *Expert Systems with Applications*, vol. 42, no. 4, pp. 2013–2025, 2015.

[18] W. Gao, L. Hu, and P. Zhang, "Class-specific mutual information variation for feature selection," *Pattern Recognition*, vol. 79, pp. 328–339, 2018.

[19] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proceedings of the 20th international conference on machine learning (ICML-03)*, 2003, pp. 856–863.

[20] X. Zhang, Z. Shi, X. Liu, and X. Li, "A hybrid feature selection algorithm for classification unbalanced data processsing," in *2018 IEEE International Conference on Smart Internet of Things (SmartIoT)*. IEEE, 2018, pp. 269–275.

[21] H. Vafaie and K. De Jong, "Genetic algorithms as a tool for feature selection in machine learning," in *Proceedings Fourth International Conference on Tools with Artificial Intelligence TAI'92*. IEEE, 1992, pp. 200–203.

[22] I.-S. Oh, J.-S. Lee, and B.-R. Moon, "Hybrid genetic algorithms for feature selection," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 26, no. 11, pp. 1424–1437, 2004.

[23] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 21, no. 12, pp. 1357–1362, 1999.

[24] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Proceedings Third IEEE international conference on automatic face and gesture recognition*. IEEE, 1998, pp. 200–205.

[25] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron," in *Proceedings Third IEEE International Conference on Automatic face and gesture recognition*. IEEE, 1998, pp. 454–459.

[26] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.

[27] K. Choi, K.-A. Toh, and H. Byun, "Realtime training on mobile devices for face recognition applications," *Pattern recognition*, vol. 44, no. 2, pp. 386–400, 2011.

[28] P. Suja, S. Tripathi *et al.*, "Real-time emotion recognition from facial images using raspberry pi ii," in *2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN)*. IEEE, 2016, pp. 666–670.

[29] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.

[30] R. L. Thorndike, "Who belongs in the family?" *Psychometrika*, vol. 18, no. 4, pp. 267–276, 1953.

[31] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.