Analysis of hand-crafted and learned feature extraction methods for real-time facial expression recognition

Jesus Olivares-Mercado, Karina Toscano-Medina,
Gabriel Sanchez-Perez, Jose Portillo-Portillo,
Hector Perez-Meana
Instituto Politecnico Nacional, ESIME Culhuacan
Mexico City, Mexico
jolivares@ipn.mx

Gibran Benitez-Garcia
Toyota Technological Institute
Nagoya, Japan
gibran@toyota-ti.ac.jp

***Abstract*—This paper presents an analysis of hand-crafted and learned feature extraction methods for real-time facial expression recognition (FER). Our analysis focuses on methods capable of running on mobile devices, including traditional algorithms such as Gabor transform, HOG, LBP, as well as two compact CNN models, named Mobilenet V1 and V2. Additionally, we test the performance of MOTIF, a highly efficient texture feature extractor algorithm. Furthermore, we analyze the contribution of the mouth and front-eyes regions for recognizing the seven basic facial expressions. Experimental results are evaluated on two publicly available datasets. KDEF database which was captured under controlled conditions and RAF database which represents more naturalistic expressions captured in-the-wild. Under the same experimental conditions, MOTIF presents the fastest performance by sacrificing accuracy, while Mobilenet V2 presents the highest results with considerable speed and model size.**

***Index Terms*—facial expression recognition, ROI, MOTIF, mobilenets, mobile applications**

I. Introduction

Facial expressions are one of the straight ways to recognize emotional states of human beings, useful to send messages to other people without words. Most of the information in human-to-human communication is transmitted by facial expressions [1], [2]. Seven basic and universal facial expressions have been defined by several psychological works [1]. The expressions of anger, disgust, fear, happiness, sadness, surprise and neutrality (no expression), are defined by the movement of different facial muscles involving facial regions such as front, eyes-eyebrows, lips, nose, and mouth [1], [2].

In recent years Facial Expression Recognition (FER) has been an increasing topic in the field of computer vision [3]–[5]. FER systems are mainly applied to human-computer interaction, behavior understanding and security approaches [2], [3]. On the other hand, the increasing use of smartphones, tablets, and mobile devices, have pushed the development of more efficient FER methods, which in most cases have to perform with low computational resources [6]. In this case, the system must operate in real-time. Thus, the complexity of the algorithms should be kept at the minimum, and efficient processes need to be implemented to run in that kind of devices smoothly [7].

The pipeline of FER as well as any biometric system can be divided in three steps: pre-processing, feature extraction, and classification [2], [4], [5]. Pre-processing is in charge of manipulating the caputerd image to facilitate the next step. Face detection, image normalization, and face segmentation are some examples of proccess used in this step [4], [6]. Feature extraction defines relevant characteristics of the face needed to represent a facial expression. Extracted features can be divided into two categories: hand-crafted features obtained with traditional methods such as Gabor transform [8], Local Binary Patterns (LBP) [9], and Histogram of Oriented Gradient (HOG) [10]; and learned features which are mostly represented by Convolutional Neural Networks (CNN) [5], [11]. Finally, the classifier generates the models needed to take the final decision. Support Vector Machine (SVM) is the most common classifier used in FER [3], [4], [7].

Detection of facial regions of interest (ROI) plays a crucial role in FER performance. Recent works demonstrate that specific facial areas contribute to a different level for expression recognition [12]–[15]. Some proposals conclude that front-eyes region contributes the most on FER [13], [14], while others state that mouth is the region which provides better results [12], [15]. The lack of studies presenting a fair comparison of different ROIs contribution could represent a problem for design robust FER systems.

In this paper, we present an analysis of FER methods known to be capable of real-time performance. We focus on traditional feature extraction algorithms including Gabor transform, LBP, HOG, and MOTIF [16] which is an efficient texture feature extractor that has been used for face recognition [17] but never tested on FER. As well as deep learning approaches, by evaluating two compact CNN models which were specifically proposed for mobile applications: Mobilenet V1 [18] and V2 [19]. Furthermore, we analyze the contribution of the mouth and front-eyes regions for recognizing the seven basic facial expression. For the ROI analysis, we employed the

KDEF database [20], a standard database which comprises 980 facial images captured in a controlled environment. The real-time analysis was evaluated using RAF dataset [21], a more naturalistic database which includes 15,339 images captured in-the-wild.

Similar works have analyzed different FER methods. However, most of them usually focus on accuracy evaluations, excluding processing speed and storage information which are crucial for implementing effective mobile applications [3]–[5], [7]. Sajjanhar et al. [11] analyze different CNN models for FER, comparing accuracy performance based on specific pre-trained models. On the other hand, Deshmukh et al. [6] present a survey of FER methods capable of working on real-time. However, they do not consider the contribution of specific facial regions.

The main contributions of this paper are two-fold: 1) An analysis of real-time FER methods evaluated under a naturalistic environment by accuracy, feature-length, inference speed and model size; 2) An analysis of the contribution of two ROI on FER, which suggests that different facial regions contribute in a different level for certain facial expressions.

II. Hand-crafted Feature Extraction Methods

In this section, we describe the four hand-crafted feature extractors used in the analysis: Gabor transform, LBP, HOG, and MOTIF. It is worth mentioning that to the best of our knowledge, this is the first time that MOTIF is tested for FER systems.

A. Gabor Transform

Gabor transform [8] is a very popular function that is used for image processing applications like face and texture recognition [8], [22] capable to support light changes. Gabor transform has frequency responses with specific orientations, frequency-selective properties, and joint optimum resolution in both spatial and frequency domains. The 2D Gabor functions are given by

$$h(y, x, i, k) = g(x', y') \exp(j2\pi F_i x') \tag{1}$$

where (x, y) express the location in the spatial domain, $F_i = \pi/2^{(i+1)}$, $i = 1, 2, \cdots, N$ F is the spatial frequency, $\Phi_k = k\pi/N_\Phi$, $k = 1, 2, \cdots, N$ Φ is the rotation angle, and $g(x', y')$ is the 2D Gaussian function given by

$$g(x', y') = \frac{1}{2\pi\sigma^2} exp\left(-\frac{x'2 + y'^2}{2\sigma^2}\right) \tag{2}$$

where $\sigma = N/2$, and N is the number of blocks in the x axis and

$$(x', y') = x\cos\Phi_k + y\sin\Phi_k - x\cos\Phi_k + y\cos\Phi_k \tag{3}$$

B. Local Binary Pattern (LBP)

Local Binary Pattern (LBP) [9] uses windows of $N \times N$ pixels, representing a neighborhood around the central pixel, as shown in Figure 1(a), where the central pixel is used to compare against its neighbors. Subsequently, if the value of the neighbor is smaller than the central, the comparison is labeled with 0, otherwise is labeled with 1, Figure 1(b). After performing all comparisons, each neighbor is multiplied by 2^P, where $0 \leq P \leq 7$ represents each pixel position in the neighborhood, Figure 1(c). Finally, all values obtained are added and the result replaces the central pixel value of the window, Figure 1(d). This process is aplied to all image pixels to produce a LBP matrix (LBP image). The original LBP algorithm uses the histogram of the final result at the end of the method.

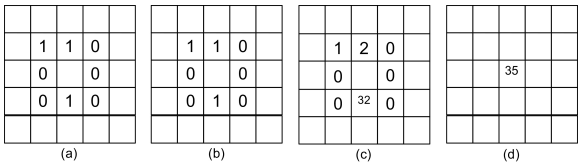

Fig. 1. LBP algorithm. (a) Values of neighbors. (b) Comparison of each neighbor with the central pixel. (c) Substitution of each value of comparison by the corresponding 2^P value. (d) Adding and replacing of central pixel with the resultant value.

C. Histogram of Oriented Gradient (HOG)

The process of Histogram of Oriented Gradient (HOG) [10] is represented in the Figure 2. To extract HOG features from a 2D image is necesary to obtain the gradient orientations of all image pixels (Figure 2(a)). Subsequently, it is necesary to get and normalize a histogram of each orientation in a small rectangular region (Figure 2(b)). Finally all the histograms that were obtained are concatenated into the final feature vector (Figure 2(c)). HOG has been commonly applied to pedestrian recognition [10], which presents a very good performance under illumination changes. Another characteristic of HOG is its robustness against deformations.

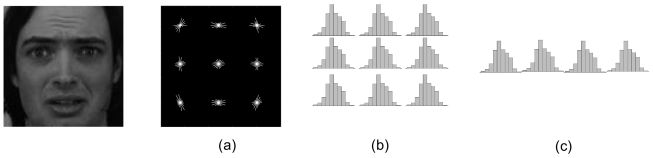

Fig. 2. HOG Algorithm: (a) gradient orientations. (b) Histogram of each region. (c) Concatenated histograms.

D. MOTIF

MOTIF algorithm was proposed for texture characterization [16]. In recent years this efficient algorithm was applied to face recognition, obtaining good performance [17]. For this reason, we propose to analyze its efficiency on FER applications. This algorithm divide full image in windows of 2×2 pixels, where

the upper left pixel in each window is consider as the starting point. Subsequently, this point is compared with the rest three pixels in the window. The pixel that has the shortest distance from the starting point is the next to be compared with the last two neighbors. The distance between pixels is given by

$$P_d = \min(P_{sp} - P_i) \qquad i = 1, 2, 3 \tag{4}$$

where P_d is the distance from the starting point (P_{sp}) and the next neighbor pixel (P_i). There exist only six MOTIF patterns that represent the comparisons of the four pixels, as shown in Figure 3. Each of these patterns is then codified with consecutive numbers from one to six in a resulting matrix. Finally, the facial image is characterized by converting the codified matrix into a row vector.

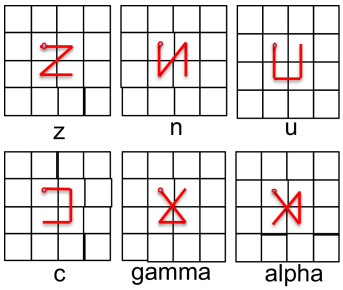

Fig. 3. MOTIF patterns.

The simplicity of this method is its main advantage. Furthermore, MOTIF is characterized to reduce the original image by $1/4$, decreasing the length of features too. As a result, training and testing time is reduced by its low computational cost. Due to these caracteristics, MOTIF is a good option to perform in mobile devices.

III. Learned Feature Extraction Methods

Thanks to the dramatically increased chip processing power of GPU units and well-designed network architecture, deep learning methods have become popular in computer vision applications [5]. Contrastively to the general trend of making deeper networks for achieving higher accuracy, Mobilenet V1 and V2 are two very recent proposals which consider efficiency concerning size and speed. In this section, we briefly describe its operation.

A. Mobilenet V1

Mobilenet V1 [18] is a novel model proposed by Google, designed specifically for mobile vision applications. The main characterisct of this proposal is the use of depthwise separable convolution layers (DSC). DSC replaces the standard convolution with a two-step operation: 1) depthwise convolution, where each $D \times D$ filter is only in charged of filtering a single depth of the input feature map; 2) pointwise convolution: a simple 1×1 pointwise convolution layer that is used for combining channel information. DSC makes the convolution operation much efficient meanwhile uses much less parameters. An example of the differences between a standard convolutions agains DSC is shown in Figure 4.

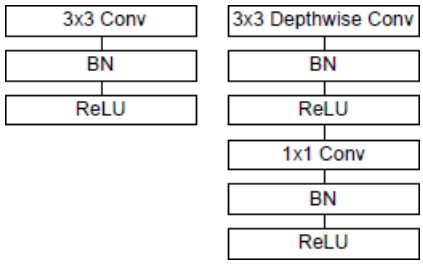

Fig. 4. Standard convolution with batchnorm and ReLU (left) against DSC with depthwise and pointwise followed by batchnorm and ReLU (right) [18].

B. Mobilenet V2

Mobilenet V2 [19] still uses DSC layers as previous version. However, V2 introduces a new module of two steps to the architecture: 1) linear bottlenecks between the layers, and 2) shortcut connections between the bottlenecks. This module expands a low-dimensional representation to high dimension and filter it with a lightweight depthwise convolution. Subsequently, features are projected back to a low-dimensional representation with a linear convolution. Finally, shortcuts enable faster training and better accuracy. The Mobilenet V2 building block is shown in Figure 5.

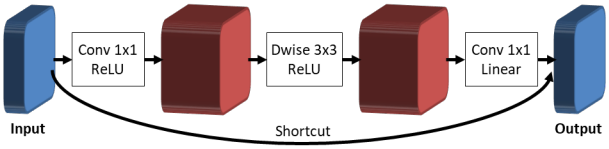

Fig. 5. Mobilenet V2 building block.

IV. Experiment Details

In this section, we describe both datasets, and present further details of the implementation of each hand-crafted and learned feature extractor as well as the SVM classifier used in our analysis.

A. Databases

Two publicly available databases were used in this paper. The KDEF database [20] which includes 980 color images of 70 people (35 female, 35 male) grouped by the seven basic facial expressions and divided in two capture sessions (each session with 490 images). This is a laboratory controlled database, which regalates illumination, position and extenral factors. Examples of each facial expressions of KDEF images are shown in Figure 6.

The Real-World Affective Faces (RAF) [21] is a large-scale database that comprises around 15,339 real-world images that were downloaded from the Internet. The whole dataset was labeled with the seven basic facial expressions, which present variability in head poses, illumination changes, occlusion, etc.

Fig. 6. Examples of KDEF images.

The database is divided by training and test sets, where the size of the training set is five times larger than the test set with expressions in equal distribution. Some images from RAF are shown in Figure 7.

Fig. 7. Examples of RAF-DB images.

B. Implementation Details

As a pre-processing step, facial regions were detected with Viola-Jones algorithm [23] and resized to 100×100 pixels. For Gabor transform we use a bank of 40 Gabor filters at five spatial scales and eight orientations. The final feature vector was set to 4,000-dimensional features by downsampling the image to 10×10. We apply the 59-bin $\text{LBP}^{u2}_{(8,2)}$ by diving the images into 100 regions with 10×10 grid size, resulting in a 5,900-dimensional feature vector. For HOG, the images were divided into 10×10 pixel blocks of four 5×5 pixel cells, obtaining a 4,000-dimensional feature vector. Since the MOTIF algorithm reduces the image into $1/4$ of the originial size, the resuling feature vector consists of 2,500-dimensional features. All hand-crafted feature vectors were classified by multi-class SVMs with RBF kernels, using the library of LIBSVM [24].

Mobilenet V1 and V2 were trained from scratch using mini-batch SGD with Nesterov momentum of 0.9, and mini-batches of 64 images. The initial learning rate was set to 0.01 and decayed by 0.98 every epoch until convergence, which required 260 and 200 epochs for Mobilenet V1 and V2 respectively. As stated in the original publications [18], [19], the dimension of feature vectors for Mobilenet V1 and V2 defined by its last fully connected layer are 1,024 and 1,280, respectively. It is important to mention that both CNN models provide an end-to-end facial expression recognition results, not used as feature extractors only.

V. Experimental Results

We present average recognition rates and confusion matrices for measuring the accuracy performance of all feature extraction methods. Tests were performed on a machine with Intel Core i5 processor at 2.40GHz, and 8GB of RAM. CNN models were trained on a single Nvidia GeForce GTX 1080.

A. Evaluation of Feature Extraction Methods for real-time FER in-the-wild

The analysis of real-time performance from hand-crafted and learned feature extraction methods was evaluated in a naturalistic environment provided by the RAF dataset. Table I shows the average recognition rate of each process. As expected, deep learning approaches (learned feature extractors) provided the best accuracy, where Mobilenet V2 obtained the highest result (81.55%). From hand-crafted feature extractors, Gabor transform presents a competitive result of 77.28%, which is about 4% lower than that of Mobilenet V2. It is worth mentioning that even the results of MOTIF are the lowest, 59% of accuracy on an unconstrained database with more than 15K images is remarkable.

TABLE I
AVERAGE RECOGNITION RATE OF DIFFERENT METHODS ON RAF DATABASE.

Method	Results (%)
Gabor	77.28
LBP	72.71
HOG	74.35
MOTIF	59.09
Mobilenet V1	80.70
Mobilenet V2	***81.55***

We present confusion matrices of the best (Mobilenet V2) and the lowest (MOTIF) results in Table II and Table III, respectively. Happiness and neutrality are the best-recognized expressions. Interestingly, even the approach of MOTIF achieved more than 80% of accuracy on happiness, 20% higher than the runner-up. In both methods, disgust was the most challenging expression to recognize, which was often misclassified as neutral. An important characteristic is that Mobilenet V2 learned in a more human-style way since fear was misclassified with surprise as a human being used to [1], [15]. On the other hand, MOTIF often misclassified fear with happiness, which is a non-natural confusion especially because of the clear differences presented in the mouth region.

TABLE II
CONFUSION MATRIX OF THE BEST RESULT ON RAF DATABASE.

	Ang	**Dis**	**Fea**	**Hap**	**Sad**	**Sur**	**Neu**
Ang	**69.75**	8.64	1.23	6.17	4.94	4.32	4.94
Dis	5	**43.13**	1.25	13.13	10.63	4.38	22.50
Fea	6.76	2.70	**50**	9.46	8.11	18.92	4.05
Hap	0.42	1.10	0.59	**92.24**	0.76	0.93	3.97
Sad	0.42	2.93	0.63	6.28	**77.41**	0.63	11.72
Sur	2.43	0.61	3.04	4.26	1.22	**80.24**	8.21
Neu	0.88	2.35	0.44	3.68	7.06	3.82	**81.76**

The efficiency results of all methods are shown in Table IV. Feats. Dim. refers to the features dimension needed to represent the whole face, the model size considers the storage required, and inference times of feature extraction and

TABLE III
CONFUSION MATRIX OF THE LOWEST RESULT ON RAF DATABASE.

	Ang	Dis	Fea	Hap	Sad	Sur	Neu
Ang	**35.19**	3.09	0.62	21.60	17.28	4.32	17.90
Dis	8.75	**11.25**	0.63	22.50	21.25	4.38	31.25
Fea	6.76	0	**27.03**	24.32	17.57	12.16	12.16
Hap	1.27	0.42	0.08	**83.88**	5.82	1.86	6.67
Sad	3.77	0.63	0.21	31.17	**34.73**	3.97	25.52
Sur	3.34	1.22	1.22	20.36	10.64	**41.95**	21.28
Neu	1.76	1.62	0.15	15.59	12.50	6.62	**61.76**

classification processes are also shown. MOTIF is the most efficient algorithm from all analyzed, it can perform at 14.6 ms (equivalent to 66 fps) and only requires 3 MB to store the SVM models. Mobilenet approaches also present a significant performance, both of them are even faster than the Gabor approach. Furthermore, if the computational power is not an issue and a GPU unit is enable, Mobilenet V1 outperforms all approaches by reaching an operation time of 5.4 ms (185 fps) almost three times faster than MOTIF. Based on the limitations of a typical mobile device (mainly CPU operation) and considering the accuracy-efficiency relation, Mobilenet V2 presents the best results with about 82% of accuracy while running at 125 ms (8 fps). Additionally, HOG still presents a competitive alternative with 74% of accuracy at 23 ms (43 fps). It is worth noting that we are not considering the computational time needed for pre-processing.

TABLE IV
EFFICIENCY RESULTS OF ALL METHODS ANALYZED IN THIS PAPER.

Method	**Feats. Dim.**	**Model Size**	**Time in miliseconds (ms)**		
			Fea. Ext.	***SVM***	***Total***
Gabor	4,000	10.7 MB	121	28.1	149.1
LBP	5,900	6.6 MB	3.4	25.2	28.6
HOG	4,000	5.9 MB	2.5	20.6	23.1
MOTIF	2,500	***3 MB***	2	12.6	***14.6***
Mobilenet V1	1,024	16.3 MB	-	-	108 (5.4*)
Mobilenet V2	1,280	8.7 MB	-	-	125 (7.3*)

*GPU time.

B. Evaluation of FER with Different ROIs

We divided each face of the KDEF dataset into two facial regions of interest (ROI). Front-eyes and mouth regions were automatically cropped based on the bounding box detection of Viola-Jones, examples of both ROI from each facial expression are shown in Figure 8. Since KDEF dataset does not explicitly provide training/testing sets, we divide the images from each subject in a ratio of 120/20, evaluated by 7-fold cross-validation. Table V presents the results of each hand-crafted feature extractor when using the whole face, front-eyes, mouth and both ROIs for its operation. For both ROIs test, we concatenated the individual feature vectors from each facial region before applying SVM.

From Table V, we can notice a significant improvement on the average recognition rate when ROI pre-processing is used. All methods improve more than 10% when features from both

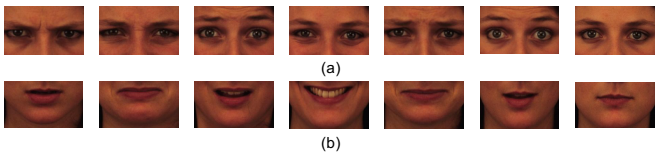

Fig. 8. Examples of ROI segmentation: (a) Front-eyes region and (b) mouth region.

ROIs are independently extracted. Particularly, MOTIF boosts its performance in more than 20% with respect to that of the whole face modality. This improvement shows that the ROI pre-processing is crucial for FER when hand-crafted feature extractors are employed, as mentioned in previous works [12]–[15]. Interestingly, MOTIF result of both ROI modality even overcomes that of Gabor using the whole face. Besides, FER performance slightly improves when front-eyes is employed rather than using only the mouth. However, this feature is not presented when LBP is used.

TABLE V
AVERAGE RECOGNITION RATE (%) OF FEATURE EXTRACTORS USING DIFFERENT ROIS.

Method	**Whole Face**	**Mouth**	**Front-Eyes**	**Both**
Gabor	***79.29***	82.50	85.89	90.71
LBP	76.43	***86.61***	***86.07***	90.36
HOG	77.86	79.82	***86.07***	***91.79***
MOTIF	60.71	68.21	77.86	84.46

We also present individual results of the highest average racognition rates from method/ROI relation. So that, Gabor/Whole-Face, LBP/mouth, HOG/front-eyes and HOG/both-ROIs results are shown in Figure 9. We can observe that each ROI contributes to a different level for recognizing individual facial expressions. For example, the front-eyes region can better recognize the expressions of disgust, surprise, and neutrality, while the mouth contributes more to anger and fear. Happiness and sadness are the most recognizable expressions for individual ROIs. Besides, the combination of both ROIs boosts their accuracy. Thus, we can justify the short average recognition difference between both ROIs presented in Table V. Interestingly, the whole face modality presents the lowest results in most cases, especially for fear and neutrality.

The confusion matrix of the highest result obtained by the ROI combination of HOG is presented in Table VI. The best results were related to happiness and sadness which achieved perfect recognition. On the other hand, even with this controlled database that presents less challenge, we found problems on recognizing disgust and anger, which are often misrecognized between them.

ROI pre-processing represents a remarkable improvement of FER accuracy with an almost imperceptive increase of latency. For instance, the HOG method which increases about 15% of accuracy, only slow down the inference speed by 2.5 ms. This low increment of computational cost can be even null if the implementation of ROI feature extraction is done in parallel.

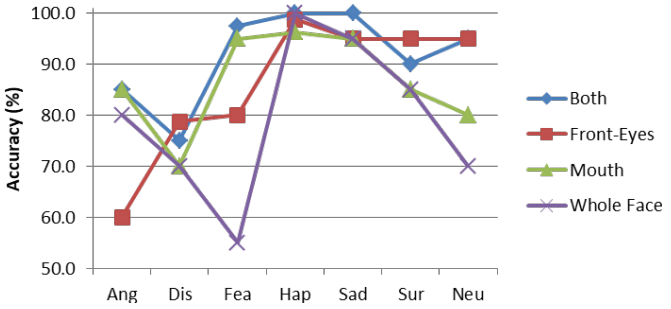

Fig. 9. Results by expression from the best average accuracy performance of each ROI.

TABLE VI
CONFUSION MATRIX OF THE BEST RESULT (HOG) USING BOTH ROIS.

	Ang	Dis	Fea	Hap	Sad	Sur	Neu
Ang	**85**	3.75	2.5	5	0	3.75	0
Dis	10	**75**	10	0	0	5	0
Fea	0	0	**97.5**	0	0	2.5	0
Hap	0	0	0	**100**	0	0	0
Sad	0	0	0	0	**100**	0	0
Sur	10	0	0	0	0	**90**	0
Neu	2.5	0	0	2.5	0	0	**95**

VI. CONCLUSIONS

In this paper, we presented an analysis of hand-crafted and learned feature extraction methods for real-time FER, as well as an analysis of the contribution of the mouth and front-eyes regions for recognizing the basic facial expressions. Experimental results showed that MOTIF is the fastest (66 fps) but presents accuracy problems (59%), while Mobilenet V2 shows the highest results (82%) with considerable speed (8 fps) and model size (8.7MB). Furthermore, front-eyes region seems to contribute slightly more than mouth for FER, and the combination of both regions presents the best results, even better than using the whole face. Additionally, we found that each region contributes to a different level based on the expression.

ACKNOWLEDGMENTS

We thank the National Science and Technology Council of Mexico and the Instituto Politecnico Nacional for the financial support during the realization of this work.

REFERENCES

[1] P. Ekman, "Facial expressions," *Handbook of cognition and emotion*, vol. 16, no. 301, p. e320, 1999. 1, 4

[2] Y. Tian, T. Kanade, and J. F. Cohn, *Facial Expression Recognition*. London: Springer London, 2011, pp. 487–519. 1

[3] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 6, pp. 1113–1133, 2015. 1, 2

[4] I. M. Revina and W. S. Emmanuel, "A survey on human face expression recognition techniques," *Journal of King Saud University-Computer and Information Sciences*, 2018. 1, 2

[5] S. Li and W. Deng, "Deep facial expression recognition: A survey," *arXiv preprint arXiv:1804.08348*, 2018. 1, 2, 3

[6] S. Deshmukh, M. Patwardhan, and A. Mahajan, "Survey on real-time facial expression recognition techniques," *IET Biometrics*, vol. 5, no. 3, pp. 155–163, 2016. 1, 2

[7] M. H. Siddiqi, M. Ali, M. E. A. Eldib, A. Khan, O. Banos, A. M. Khan, S. Lee, and H. Choo, "Evaluating real-life performance of the state-of-the-art in facial expression recognition using a novel youtube-based datasets," *Multimedia Tools and Applications*, vol. 77, no. 1, pp. 917–937, 2018. 1, 2

[8] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Transactions on Image processing*, vol. 11, no. 4, pp. 467–476, 2002. 1, 2

[9] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002. 1, 2

[10] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 886–893. 1, 2

[11] A. Sajjanhar, Z. Wu, and Q. Wen, "Deep learning models for facial expression recognition," in *2018 Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 2018, pp. 1–6. 1, 2

[12] G. Benitez-Garcia, G. Sanchez-Perez, H. Perez-Meana, K. Takahashi, and M. Kaneko, "Facial expression recognition based on facial region segmentation and modal value approach," *IEICE TRANSACTIONS on Information and Systems*, vol. 97, no. 4, pp. 928–935, 2014. 1, 5

[13] A. Hernandez-Matamoros, A. Bonarini, E. Escamilla-Hernandez, M. Nakano-Miyatake, and H. Perez-Meana, "Facial expression recognition with automatic segmentation of face regions using a fuzzy based classification approach," *Knowledge-Based Systems*, vol. 110, pp. 1–14, 2016. 1, 5

[14] D. Ghimire, S. Jeong, J. Lee, and S. H. Park, "Facial expression recognition based on local region specific features and support vector machines," *Multimedia Tools and Applications*, vol. 76, no. 6, pp. 7803–7821, 2017. 1, 5

[15] G. Benitez-Garcia, T. Nakamura, and M. Kaneko, "Multicultural facial expression recognition based on differences of western-caucasian and east-asian facial expressions of emotions," *IEICE TRANSACTIONS on Information and Systems*, vol. 101, no. 5, pp. 1317–1324, 2018. 1, 4, 5

[16] A. Obulesu, V. V. Kumar, and L. Sumalatha, "Content based image retrieval using multi motif co-occurrence matrix." *International Journal of Image, Graphics & Signal Processing*, vol. 10, no. 4, 2018. 1, 2

[17] J. Olivares-Mercado, K. Toscano-Medina, G. Sanchez-Perez, M. Nakano-Miyatake, and H. Perez-Meana, "Face recognition system based on motif features," *Journal of Modern Optics*, vol. 65, no. 18, pp. 2124–2132, 2018. 1, 2

[18] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017. 1, 3, 4

[19] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510–4520. 1, 3, 4

[20] D. Lundqvist, A. Flykt, and A. Öhman, "The karolinska directed emotional faces (kdef)," *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet*, vol. 91, p. 630, 1998. 2, 3

[21] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 2584–2593. 2, 3

[22] M. Pakdel and F. Tajeripour, "Texture classification using optimal gabor filters," in *Computer and Knowledge Engineering (ICCKE), 2011 1st International eConference on*. IEEE, 2011, pp. 208–213. 2

[23] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–I. 4

[24] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. 4