PAPER Special Section on Machine Vision and its Applications

Multicultural Facial Expression Recognition Based on Differences of Western-Caucasian and East-Asian Facial Expressions of Emotions

Gibran BENITEZ-GARCIA[†a)], **Tomoaki NAKAMURA**[†], ***Nonmembers***, ***and*** **Masahide KANEKO**[†], ***Fellow***

SUMMARY An increasing number of psychological studies have demonstrated that the six basic expressions of emotions are not culturally universal. However, automatic facial expression recognition (FER) systems disregard these findings and assume that facial expressions are universally expressed and recognized across different cultures. Therefore, this paper presents an analysis of Western-Caucasian and East-Asian facial expressions of emotions based on visual representations and cross-cultural FER. The visual analysis builds on the Eigenfaces method, and the cross-cultural FER combines appearance and geometric features by extracting Local Fourier Coefficients (LFC) and Facial Fourier Descriptors (FFD) respectively. Furthermore, two possible solutions for FER under multicultural environments are proposed. These are based on an early race detection, and independent models for culture-specific facial expressions found by the analysis evaluation. HSV color quantization combined with LFC and FFD compose the feature extraction for race detection, whereas culture-independent models of anger, disgust and fear are analyzed for the second solution. All tests were performed using Support Vector Machines (SVM) for classification and evaluated using five standard databases. Experimental results show that both solutions overcome the accuracy of FER systems under multicultural environments. However, the approach which individually considers the culture-specific facial expressions achieved the highest recognition rate.

key words: *facial expression recognition, multicultural FER, culture specificity of facial expressions of emotions, universality of emotions*

1. Introduction

Facial expressions are a set of facial muscle movements which can directly express human emotions. Charles Darwin was the first one who tried to reveal the origins of facial expressions [1]. He claimed that facial expressions are innate and evolved human behaviors, which can be recognized across different races and cultures around the world. It is worth noting that facial expressions are part of the communication process among humans, which involves the signaling and decoding of information (facial expression). In order to measure both sides of the process, Paul Ekman et al. [2] proposed the Facial Action Coding System (FACS) which is focused on 44 anatomical facial muscle movements called Action Units (AUs), with this system they standardized the prototypic expressions of anger, disgust, fear, happiness, sadness and surprise [3]. Thus, the universality of six basic facial expressions of emotions was established.

Contrastively, some researchers argued that the conclusions of those studies do not consider the misclassification errors which can be affected by cultural differences. I.e., in-group advantages which define that each person tends to appreciate other people's facial expressions based on their own cultural knowledge [4]. In addition, recent psychological studies have addressed the origins of cultural differences in facial expression recognition, showing that some cultures have systematic confusions on distinguishing certain expressions [5].

In spite of the increasing debate about the universality of facial expressions of emotion, from the viewpoint of the human-computer interaction (HCI), the cultural universality of emotions is taken for granted [6]. In other words, automatic facial expression recognition systems (FER) builds on the assumption that the six basic expressions of emotions are equally expressed all across different cultures and should be universally recognized. However, in order to attain a robust FER systems, the psychological findings of cultural differences of some expressions should be considered.

This paper presents an extensive analysis of the differences between Western-Caucasian and East-Asian basic facial expressions of emotions, as well as two possible solutions for overcoming the problem of multicultural FER. The analysis consists of visual representations of the main facial features extracted by the well-known Eigenfaces approach (focused on the six basic expressions from each cultural group) and cross-cultural FER based on the combination of Local Fourier Coefficients (LFC) and Facial Fourier Descriptors (FFD) which describes appearance and geometric facial features from peak expressions, respectively.

Our analysis can be seen as an extension of that proposed in [8], as well as an enhancement of the study presented in [9]. The main difference is that, in this paper, the analysis was applied to a more generalized database (more than 5 times bigger than those of [8] and [9]) which includes more than 1,000 expressive faces from five standard databases (CK+, MUG, JAFFE, JACFE and TFEID). Moreover, the feature extraction process employed for the cross-cultural FER analysis is fully automatic as proposed in [7].

In addition, this paper presents two possible solutions for the multicultural problem of FER. The first one considers a pre-step of race recognition, where the main proposal is to combine facial features of color (HSV color quantization),

Manuscript received September 7, 2017.
Manuscript revised November 29, 2017.
Manuscript publicized February 16, 2018.

[†]The authors are with Department of Mechanical Engineering and Intelligent Systems, The University of Electro-Communications, Chofu-shi, 182–8585 Japan.
a) E-mail: gibran@toyota-ti.ac.jp

DOI: 10.1587/transinf.2017MVP0025

texture (LFC) and shape (FFD) for discriminating individual's race. The second approach is based on the culture-specific facial expressions found by our analysis. These expressions are particularly handled by the classification process (training mode), where individual models are calculated for culture-specific expressions whereas multicultural models are obtained for those proven to be cross-culturally well recognized.

In summary, the main contributions of this paper include: 1) a qualitative and quantitative cross-cultural analysis of FER; 2) an early automatic detection of Western-Caucasian and East-Asian subjects as a pre-processing step for multicultural FER; 3) a proposal considering culture-specific and multicultural expressions for classifying the six basic expressions of emotion.

The rest of the paper is organized as follows. Section 2 discusses related works. The analysis framework is explained in Sect. 3 followed by the proposed multicultural FER solutions which are described in Sect. 4. The experimental results are shown in Sect. 5. Section 6 presents a discussion about solutions and findings of the analysis, and finally, conclusion and future works are drawn in Sect. 7.

2. Related Works

Related works are clearly divided by psychological and HCI viewpoints. Psychological studies try to prove the refutation of the universality hypothesis of facial expressions. Dailey et al. [10] evaluated the effect of culture-specific facial expression interpretation by analyzing the recognition capability of U.S. and Japanese participants. Their work is based on a human study using a cross-cultural dataset, and a replication of the studied human behavior by using a model based on Gabor, PCA and artificial neural networks. This experiment helps to demonstrate how the interaction with other people in a cultural context defines the way of recognizing a culture-specific facial expression dialect.

Jack et al. [11] claimed to refute the universal hypothesis of facial expressions of emotions by using generative grammars and visual perception for analyzing the mental representations of Westerns and East-Asians. Facial expression representations per culture, based on the six basic emotions were modeled and they found that each emotion is not expressed using a combination of facial movements common to both racial groups. Thus, it is demonstrated that the basic emotions can clearly represent the Western facial expressions, but those are inadequate to accurately represent East-Asian emotions, demonstrating a culture-specific based representation of the basic emotions.

On the other hand, from the HCI viewpoint, many FER systems strongly follow the universality hypothesis of facial expressions and perform cross-database evaluations in order to only prove their robustness. However, a few studies attend the cultural differences of the available databases. For example, Da Silva and Pedrini [12] presented a cross-cultural FER analysis using occidental and oriental face databases. The analysis was based on 3 different standard feature extraction methods and 3 machine learning algorithms. The best results obtained were achieved by the in-group test, followed by those of the multicultural test. However, when the out-group test was applied, the accuracy dramatically decreased. The authors concluded that multicultural training should be considered when an efficient recognition performance is needed.

Ali et al. [13] performed a similar study, where ensemble classifier construction was intended to find how the classifiers will be trained to accurately classify multicultural databases. This proposal utilized boosted NNE (Neural Network Ensemble), HOG features and Naïve Bayes for cross-classifying facial expressions from Moroccan and Caucasian subjects with those from Japan and Taiwan. The results reported follow the same trend as [12] (in-group > multicultural > out-group). Therefore, the authors concluded that promising results are obtained when multicultural databases are used for training. In addition, they attached the problems of out-group performance to the inconsistency in the number of samples per expression, visual representation and facial structure.

In general, most of the automatic FER studies attribute the multicultural and out-group problems to external factors such as algorithm robustness or image quality rather than questioning the universality of facial expressions itself.

3. Analysis Framework

3.1 Datasets

The analysis is evaluated using a total of 1,200 facial images from 254 subjects, which were selected from five standard datasets. The complete whole set, from now called multicultural dataset (MUL), was divided into two racial groups: Western-Caucasian (WSN) and East-Asian (ASN) dataset. WSN dataset comprises 600 expressive images (100 per basic expression) from 149 Western subjects selected from the extended Cohn-Kanade dataset (CK+) [14] and the Multimedia Understanding Group Facial Expression database (MUG) [15]. In turn, ASN contains the same number of images from 105 East-Asians taken from Japanese Female Facial Expression dataset (JAFFE) [16], Japanese and Caucasian Facial Expression of Emotion dataset (JACFEE) [17], and the Taiwanese Facial Expression Image Database (TFEID) [18]. Figure 1 illustrates some faces of the six basic expressions included in both datasets. Not shown are the images selected from JACFFE (Japanese only) which cannot be reprinted due to copyright restrictions.

3.2 Visual Analysis

The visual representation of facial features is based on the well-known algorithm of Eigenfaces. Reconstructed images from feature vectors projected into the previously calculated Eigenspace give the opportunity to analyze possible differences among the basic expressions of each cultural group.

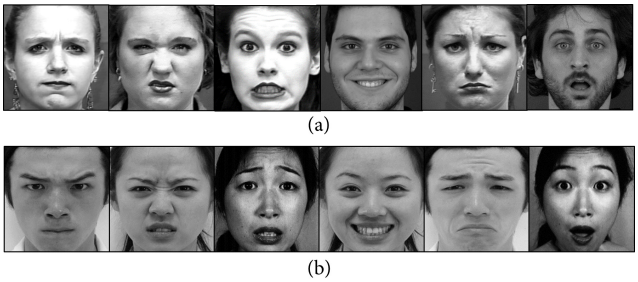

Fig. 1 Images included in both datasets: (a) WSN, (b) WSN. From left to right: anger, disgust, fear, happiness, sadness and surprise.

Thus, as proposed in [8], we obtain average projected vectors by each expression, which can be calculated from a cultural-specific dataset by:

$$\boldsymbol{Z} = \frac{1}{P}\sum_{i=1}^{P}\boldsymbol{Y}(i) \tag{1}$$

where $\boldsymbol{Z}$ represents the average projected vector, P the number of images of the current expression (for this paper $P = 100$), and $\boldsymbol{Y}$ the feature vector of each facial image. Finally, reconstructed images are the reshaped matrix of reconstructed average projected vectors defined by:

$$\boldsymbol{R} = \boldsymbol{\Phi Z} + \boldsymbol{\mu} \tag{2}$$

where $\boldsymbol{\Phi}$ is the Eigenspace of all facial images, and $\boldsymbol{\mu}$ represents the mean of feature vectors.

Figure 2 shows the visual representation of the average expressions of both datasets. Similar to the findings of [8] and [9], we can clearly observe that disgust and fear expressions look different. Disgusted average face from WSN presents the common AUs for disgust (AU9, AU15, AU16). However, the same average face from ASN shows differences in the eyes region, specifically AU22 and AU23, which are known to appear in a common anger expression. The average WSN face of fear includes all the emotion FACS for a common fear expression (AU1, AU2, AU4, AU5, AU7, AU20, AU26), whereas that of ASN lacks of AU4 and AU20, given the impression of surprised eyes. The mouth region of average anger of ASN looks similar to that of WSN, still, it seems to have many variations within individual faces. On the other hand, expressions of happiness, sadness and surprise do not present significant visual differences among cultures.

3.3 Cross-Cultural FER

The FER system used for cross-cultural evaluation is based on the hybrid method originally proposed in [8] which was enhanced by the application of the Fast Fourier Transform (FFT), as detailed in [7]. The system framework follows the steps of face detection, facial region segmentation, feature extraction and classification. Face detection was carried out by Viola-Jones algorithm and facial landmark extraction is based on a deformable face tracking model which locates

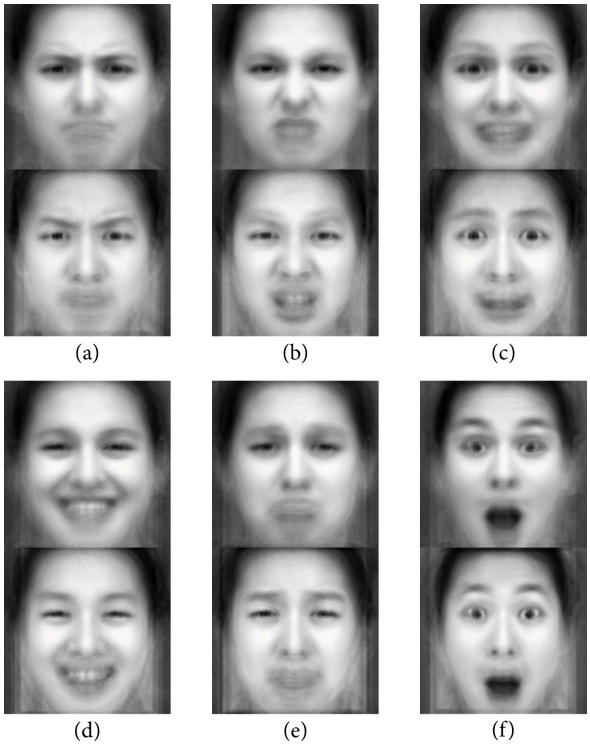

Fig. 2 Comparison among WSN (top) and ASN (bottom) average expressions. (a) Anger, (b) Disgust, (c) Fear, (d) Happiness, (e) Sadness, (f) Surprise.

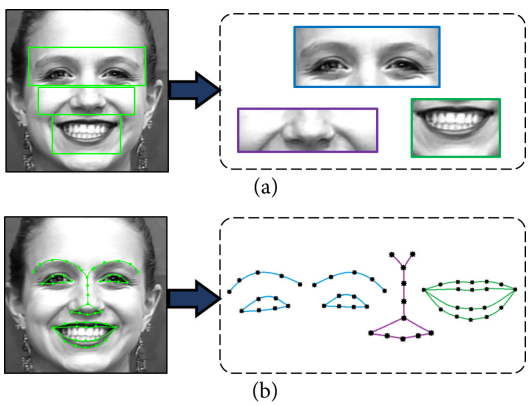

Fig. 3 Example of facial region segmentation as proposed in [7]. (a) Appearance and (b) geometric features.

51 facial points for describing the whole shape of the face. Facial region segmentation is based on the distance between irises and it is applied for both kind of features, so that, appearance and geometric information of eyes-eyebrows (from now called "eyes" for simplicity), nose and mouth are individually obtained. Figure 3 shows an example of these facial regions.

Feature extraction process generates hybrid feature vectors obtained from the fusion of Local Fourier Coefficients (LFC) and Facial Fourier Descriptors (FFD). Basically, LFC and FFD are based on the application of the FFT and the calculation of individual eigenspaces for each fa-

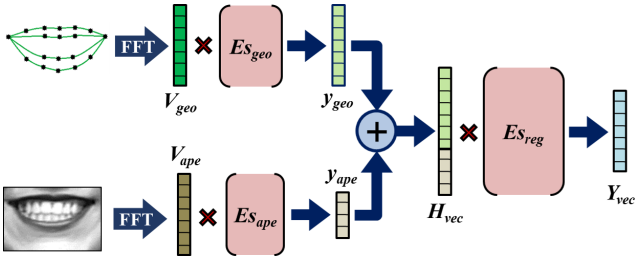

Fig. 4 Example of the feature extraction process, as proposed in [7].

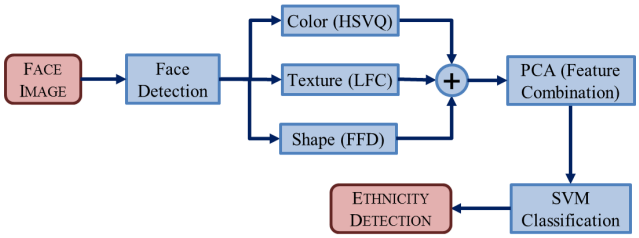

Fig. 5 Race detection framework.

cial part, which are used to obtain independent feature vectors of appearance and geometric features, respectively. Hybrid feature vectors are obtained by concatenating independent feature vectors, which in turn are projected into a final region-based eigenspace (for details of this method, please refer to [7]). Figure 4 illustrates the feature extraction process of the mouth region, where $\boldsymbol{Y}_{vec}$ represents the hybrid feature vector, which is a projection of the concatenated $\boldsymbol{H}_{vec}$ vector into the $\boldsymbol{Es}_{reg}$ mouth Eigenspace. Similarly, $\boldsymbol{y}_{geo}$ and $\boldsymbol{y}_{ape}$ represent individual feature vectors which are projections of $\boldsymbol{V}_{geo}$ and $\boldsymbol{V}_{ape}$ vectors into the Eigenspaces of geometric ($\boldsymbol{Es}_{geo}$) and appearance ($\boldsymbol{Es}_{ape}$) features. It is worth noting that $\boldsymbol{V}$ vectors are obtained after the application of FFT.

Finally, the classification stage was independently performed by SVM based on different cross-cultural recognition modalities: in-group, out-group and multicultural. In-group classification represents FER performance when the same cultural-specific dataset is used for training and testing. Out-group classification presents the opposite situation, when training phase is carried out with a different dataset of that used for testing. Multicultural classification occurs when training and testing are conducted using a multicultural dataset.

4. Multicultural Solutions

4.1 Early Race Detection

This proposal builds its process on a logical solution, an early race detection of WSN and ASN subjects. Thus, the FER system has to be trained separately for race. In turn, the testing phase is performed employing the models which correspond to the previously detected race of the input face. In other words, the pre-processing stage of race detection decides the cultural sub-space where the facial expression will be evaluated.

The race detection process is based on three different features obtained from the detected face in the input image. These are texture, shape and color features, where LFC and FFD processes are employed for extracting texture and shape features, while color feature extraction is based on a proposed modification of the Dominant Color Correlogram Descriptor (DCCD), a method which is used for content-based image retrieval [19]. After obtaining the three different features, these are combined by Principal Component Analysis (PCA) and classified by SVM. The process of this sub-system is shown in Fig. 5.

For color feature extraction, the whole face region should be transformed to HSV color space, subsequently, HSV is quantified so that only 72 different skin colors are considered. Finally, the histogram calculated from all the skin colors of the face region, is taken as a color feature vector. HSV Quantization (HSVQ) is based on the fact that all possible skin tones of different human races are part of a subset color from elements of HSV [20]. Many studies have found that the skin tone related to the hue component of HSV falls into to sub-group of $300^o \leq \mathrm{H} \leq 60^o$ [21]. Therefore, non-interval quantization of the HSV color space is employed, where H components are divided into eight shares and only the sub-group which includes skin colors is taken into account, this process is defined by:

$$H = \begin{cases} 0 & \text{if } 280 < h \leq 300, \\ 1 & \text{if } 300 < h \leq 320, \\ 2 & \text{if } 320 < h \leq 340, \\ 3 & \text{if } 340 < h \leq 360, \\ 4 & \text{if } 0 < h \leq 20, \\ 5 & \text{if } 20 < h \leq 40, \\ 6 & \text{if } 40 < h \leq 60, \\ 7 & \text{if } 60 < h \leq 80. \end{cases} \tag{3}$$

where h is the value of hue component of a certain pixel of the face region, and H is the new quantized value. Subsequently, S and V components are divided into three shares respectively, as given by:

$$S = \begin{cases} 0 & \text{if } 0 < s \leq \frac{1}{3}, \\ 1 & \text{if } \frac{1}{3} < s \leq \frac{2}{3}, \\ 2 & \text{if } \frac{2}{3} < s \leq 1. \end{cases} \quad V = \begin{cases} 0 & \text{if } 0 < v \leq \frac{1}{3}, \\ 1 & \text{if } \frac{1}{3} < v \leq \frac{2}{3}, \\ 2 & \text{if } \frac{2}{3} < v \leq 1. \end{cases} \tag{4}$$

where s and v represent values of saturation and "value" (from HSV hexcone model) respectively, whereas S and V their quantized values. The final step of the quantization is to obtain the combination of the three individual values, which is defined by:

$$C = 9H + 3S + V \tag{5}$$

where C represents one of the 72 possible colors of human skin.

Finally, from the new matrix obtained by all evaluated

pixels, the histogram which calculates the most recurrent colors from the face image is used as a color feature vector. Figure 6 shows two examples of this process, on the left side, we can see the original facial images, followed by the visual representation of the quantitated colors obtained by Eq. 5 (up to 72 different colors). Histograms of the color representation are shown on the right side. From these examples, we can see that the variation of the color histograms from each racial group is clearly distinctive. Therefore, these can be employed as feature vectors.

4.2 Consideration of Culture-Specific Expressions

Based on facial expressions that present cultural differences from the visual analysis of average expressions, this proposal considers different variations for training the system. The traditional way for training any FER system aims to obtain models related to each of the six basic expressions. Thus, only six models are obtained at the end of the training phase. These models depend on the dataset used for training, so that, only one cultural dataset is used for in-group and out-group modality (WSN or ASN) and the combination of both datasets are used for multicultural (MUL). In this way, the multicultural training disregards the differences that may appear among facial expressions from different cultures. However, from the psychological literature and our visual analysis, we know that there are some facial expressions that are not equally expressed among the cultures. Therefore, in this approach, the expressions of anger, disgust and fear are considered as culture-dependent. Thus, two sub-models are obtained for each of these expressions.

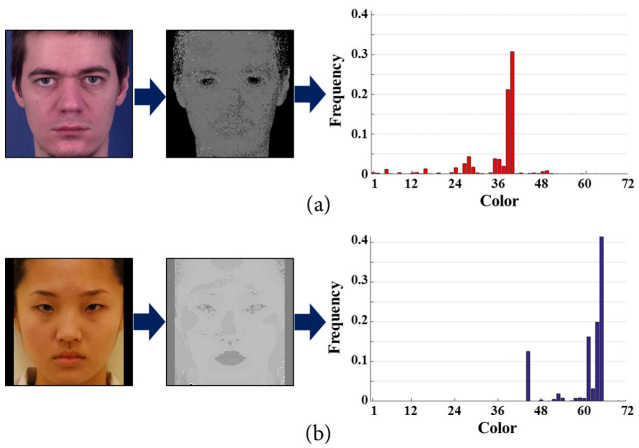

Fig. 6 Examples of the quantitated colors and histograms obtained by the HSVQ process. Applied to (a) WSN and (b) ASN faces.

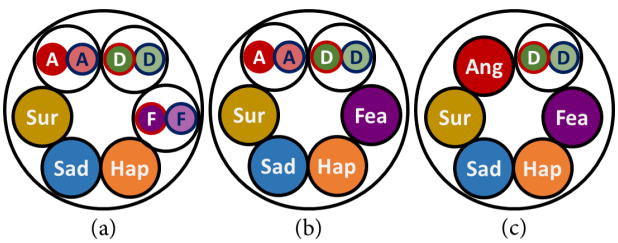

Fig. 7 Example of the possible training variations based on the culture-specific facial expressions. (a) Considering three culture-dependent models, (b) considering two, and (c) considering just one.

Consider the three diagrams illustrated in Fig. 7, each of them represents a training variation based on the previously mentioned culture-specific expressions. For instance, (a) obtains culture-specific sub-models for anger, disgust and fear (trained with WSN and ASN datasets independently) and multicultural models for happiness, sadness and surprise (trained with MUL dataset). (b) presents a similar training process but in this case, fear is considered as a multicultural model. Finally, (c) represents the variation when only disgust is independently trained based on WSN and ASN datasets. It is worth noting that the final decision of the testing phase is limited to the six basic expressions, where the race of the input face is not required.

5. Experimental Results

The six basic expressions of all tests of this paper were classified by multi-class SVMs with RBF kernels and evaluated using leave-one-subject-out (LOSO) with cross-validation. Average recognition rates and confusion matrices are used for measuring the accuracy performance.

5.1 Cross-Cultural Analysis

Table 1 shows the results of all classification modalities (in-group, out-group and multicultural) presented by individual facial region and its combinations. From this table, we can notice that in general the best results are obtained by the in-group modality, where WSN test presents the highest accuracy for most of the facial region combinations. An interesting result is that the multicultural test (MUL) presents better results than ASN but lower than WSN. Finally, the worst accuracy performance is obtained by the out-group modality, where the lowest results are presented when training with the ASN dataset. Furthermore, following logical results, the system achieves higher recognition accuracy when the feature vector combination of all facial regions (Eyes-Nose-Mouth) is employed. However, it seems that the mouth region represents an important feature for FER because the best results are obtained when it is used in the feature extraction process.

Table 2 delves into the results of all classification

Table 1 Average recognition rate (%) of all classification modalities, divided by individual facial regions and its combinations.

Training:	WSN	ASN	WSN	ASN	MUL
Testing:	WSN	ASN	ASN	WSN	MUL
Eyes	70	67.3	48.3	49.7	72.9
Nose	66.3	52.8	49.3	42.7	70.1
Mouth	85.8	77	69.3	61.3	80.3
Eyes-Nose	80.8	72.3	60.5	50.3	80.1
Eyes-Mouth	93.8	86.7	75.8	68	90.4
Nose-Mouth	88.2	81	75.5	67	86.3
All	96	85.8	79.8	68.3	90.8

Table 2 Recognition rate (%) per expression of all classification modalities using the combination of all facial regions.

Training:	WSN	ASN	WSN	ASN	MUL
Testing:	WSN	ASN	ASN	WSN	MUL
Ang	95	82	65	60	89.5
Dis	98	81	68	58	91
Fea	95	85	71	75	88.5
Hap	100	92	100	69	94.5
Sad	91	80	77	73	87
Sur	97	95	98	75	94.5

Table 3 Confusion matrix of the multicultural modality using the combination of all facial regions.

	Ang	Dis	Fea	Hap	Sad	Sur
Ang	**89.5**	1	2	0.5	7	0
Dis	2.5	**91**	2.5	1	2	1
Fea	1.5	0.5	**88.5**	3.5	4.5	1.5
Hap	0	1.5	2.5	**94.5**	1.5	0
Sad	8	0.5	4.5	0	**87**	0
Sur	0	1.5	4	0	0	**94.5**

Table 4 Average recognition rate (%) of the best multicultural solutions, divided by individual facial regions and its combinations.

Training:	MUL	Race	AnDiFe	AnDi	Di	Fe
Eyes	72.9	69.3	**75.8**	73.4	73.8	75.2
Nose	70.1	60.1	68.8	70.9	**72**	69.1
Mouth	80.3	82.2	81.3	81	81.6	**82.8**
Eyes-Nose	80.1	77.3	**81.2**	80.8	80.1	80.6
Eyes-Mouth	90.4	91.1	92.6	**93.2**	92.4	91.9
Nose-Mouth	86.3	85.4	86.9	**87.8**	87.1	87.1
All	90.8	91.8	92.9	**93.3**	92.4	92.3

modalities (using the combination of all facial regions) and presents the recognition accuracy by each facial expression. From this table we can see that WSN models are highly capable for recognizing the expression of happiness, however, these results decrease when MUL dataset is used for training. In fact, none of the facial expressions from multicultural modality overcome the results of those of WSN. Moreover, the results of anger, disgust and fear present the most significant drop in accuracy, as expected.

In order to analyze the recognition errors of multicultural test, its confusion matrix is presented in Table 3. Here we can see that a signinifcant problem is related with the misrecognition between sadness and anger. It seems that the system has problems to discriminate among these expressions. Indeed, sadness also presents relevant misrecognition problems with fear, being this the worst recognized expression.

5.2 Multicultural Solutions

Results of multicultural solutions are presented in Table 4. "MUL" refers to the multicultural test and "Race" to the solution described in Sect. 4.1. The best results of all combinations of the solution described in Sect. 4.2 are named "AnDiFe", "AnDi", "Di" and "Fe", which refer to the tests when the named expressions are treated as culture-specific sub-models. For example, "AnDiFe" refers to the training modality where anger, disgust and fear were especially considered. From Table 4 we can see that the results of individual facial regions of MUL test are easily overcome by the solutions based on the consideration of culture-specific facial expressions, this may be related to the expressive differences of specific facial regions among the cultures. Finally, the best results are obtained by the "AnDi" solution, where the expressions of anger and disgust are considered as culture-specific.

Table 5 Recognition rate (%) per expression of the best solutions for multicultural environments using the combination of all facial regions.

Training:	WSN	ASN	MUL	Race	AnDiFe	AnDi
Ang	95	82	89.5	90	89	89
Dis	98	81	91	90.5	92.5	93
Fea	95	85	88.5	90	89	93.5
Hap	100	92	94.5	96.5	97	96
Sad	91	80	87	87.5	92	91
Sur	97	95	94.5	96	98	97.5

Table 6 Confusion matrix of the AnDi multicultural solution using the combination of all facial regions.

	Ang	Dis	Fea	Hap	Sad	Sur
Ang	**89**	1.5	1	0	8.5	0
Dis	2	**93**	1	1	2	1
Fea	1	0.5	**93.5**	1.5	2.5	1
Hap	0	1	1.5	**96**	1	0.5
Sad	5	0.5	3.5	0	**91**	0
Sur	0	0	2	0	0.5	**97.5**

In order to have an easier way to compare the recognition rate per expression among different tests and solutions, Table 5 presents the results of in-group and multicultural modalities, altogether with the best results of both proposed solutions. It is possible to notice that the solutions based on the consideration of culture-specific expressions highly overcome the results of ASN test and are close to those of WSN. For example, "AnDi" test presents a considerable improvement of accuracy for the expression of fear, highly overcoming the results of ASN and MUL. In addition, results of sadness and surprise of "AnDiFe" even overcome that of WSN. Interestingly, the accuracy of anger remains low for all the possible solutions.

Finally, in order to analyze the possible misrecognition problems of the best solution, Table 6 presents the confusion matrix of "AnDi". From this table, we can see that the problems of misrecognition are still related to the expressions of anger and sadness, where anger is misrecognized with sadness in 8.5% of the cases.

6. Discussion

Thanks to the visual analysis we could confirm that the differences among both cultural groups reside on the expressions of disgust and fear, specifically for the facial regions of mouth and eyes-eyebrows, respectively (same as using smaller datasets as found in [8] and [9]). These differences are significant for the automatic FER system and can be noticed by the low recognition performance related to these

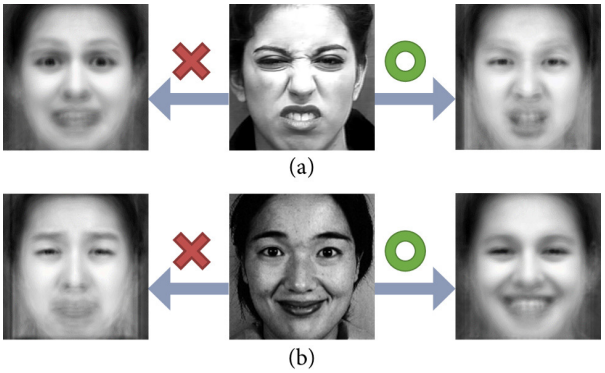

Fig. 8 False positive examples of race detection which were correctly recognized for the FER multicultural solution. (a) WSN disgust, (b) ASN happiness.

expressions.

Table 1 shows that the multicultural test (MUL) overcomes the results obtained by ASN test. However, its performance is still far from the results obtained by WSN. Moreover, the proposed multicultural solutions presented in this paper overcome the results of the multicultural test for all possible combinations of facial regions (Table 4). Interestingly, the results of the race detection approach ("Race") even overcome those of the averaged performance among WSN and ASN of the in-group test. This issue is strongly related to the cross-cultural basic expressions and it happens because some faces that were misrecognized in the race detection, were correctly recognized by FER models of the opposite cultural group. Figure 8 illustrates two examples of this situation, both subjects were misrecognized in the race detection. Figure 8 (a) shows a WSN subject displaying disgust, this sample was misrecognized with fear by the WSN training, but it was correctly recognized by the ASN training. The opposite situation occurs in Fig. 8 (b) where the ASN expression of happiness was misrecognized with sadness by the ASN training but correctly recognized by WSN. It is worth noting that the recognition accuracy of the race detection was 99.1% and all misrecognition errors were related to the databases that do not include color images (CK+ and JAFFE). The accuracy recognition of the proposed method including color features was 100%.

As mentioned before, most of the psychological studies that aim to clarify the origins of facial expression differences, found that these may be related to cultural context. Taking into account this hypothesis, the implementation of the proposed early race detection should be conscientiously considered by its application. Because it will present several misrecognition problems if the system is tested with subjects that raised on countries which are different from their ethnic. Therefore, the best solution, in this case, is the consideration of culture-specific facial expressions. Thus, as shown by our second proposal, the final decision for multicultural environments will be made based on culture-specific plus cross-cultural basic expressions regardless the ethnicity of the subject.

It is important to clarify that the analysis presented in this paper is based only on peak expressions (apex of a common expression sequence) and it is affected by the inherent problems of the use of standard datasets taken under controlled environments. Thus, a further analysis of the effect of temporal information and the treatment of FER in-the-wild may help for supporting our results. However, despite the mentioned limitations, the present study highlights differences in automatic FER of basic expressions that were believed to be universally recognized from an HCI perspective. Furthermore, it proposes the consideration of these differences for improving the performance of multicultural FER.

7. Conclusion and Future Work

In this paper, we presented an analysis of Western-Caucasian and East-Asian facial expressions of emotions. Based on the experimental results, we can conclude that similar to the psychological findings, the proposed FER system presents in-group advantages, where WSN facial expressions are easier to recognize than those of ASN. In addition, the quantitative and qualitative analysis shows that disgust and fear are culture-specific expressions, whereas the rest of the six basic expressions are cross-culturally recognized. Furthermore, the two proposed solutions for multicultural FER overcome the results of traditional multicultural approaches. However, we can conclude that the consideration of culture-specific facial expressions may be a better solution, not only because of its high accuracy but also because it can correctly classify expressions of subjects with a cultural basis different than their ethnicity.

As a future work, we are planning to improve the solution for multicultural environments by considering the facial region differences. In addition, in order to overcome the problems of databases reliability and the consideration of temporal information, we would like to analyze the culture-specific differences "in-the-wild" by employing robust algorithms which take advantage of semantic and dynamic features, such as the combination of convolutional and recurrent neural networks (CNN + RNN). In this way, we could analyze multicultural FER on real-life applications, which may handle all sequence states, such as, peak, strong, weak and null expressions.

References

[1] C. Darwin, "The expression of the emotions in man and animals," University of Chicago press, Chicago, 1965.
[2] P. Ekman and W.V. Friesen, "Facial action coding system," Consulting Psychologists Press, Palo Alto, 1977.
[3] P. Ekman, "An argument for basic emotions," Cogn. Emotion, vol.6, no.3-4 pp.169–200, 1992.
[4] H.A. Elfenbein and N. Ambady, "On the universality and cultural specificity of emotion recognition: a meta-analysis," Psychological bulletin, vol.128, no.2, pp.203–235, 2002.
[5] R.E. Jack, "Culture and facial expressions of emotion," Visual Cognition, vol.21, no.9-10, pp.1248–1286, Sept. 2013.

[6] Y. Tian, T. Kanade, and J.F. Cohn, "Facial Expression Recognition," Handbook of Face Recognition, eds. S.Z. Li and A.K. Jain, pp.487–519, Springer, London, 2011.
[7] G. Benitez-Garcia, T. Nakamura, and M. Kaneko, "Facial Expression Recognition Based on Local Fourier Coefficients and Facial Fourier Descriptors," Journal of Signal and Information Processing, vol.8, no.3, pp.132–151, 2017.
[8] G. Benitez-Garcia, T. Nakamura, and M. Kaneko, "Analysis of in- and out-group differences between Western and East-Asian facial expression recognition," Proc. 15th IAPR Int. Conf. Machine Vision Applications, Nagoya, Japan, pp.402–405, May 2017.
[9] G. Benitez-Garcia, T. Nakamura, and M. Kaneko, Methodical Analysis of Western-Caucasian and East-Asian Basic Facial Expressions of Emotions Based on Specific Facial Regions," Journal of Signal and Information Processing, vol.8, no.2, pp.78–98, 2017.
[10] R.E. Jack, O.G.B. Garrod, H. Yu, R. Caldara, and P.G. Schyns, "Facial expressions of emotion are not culturally universal," Proc. National Academy of Sciences, vol.109, no.19, pp.7241–7244, May 2012.
[11] M.N. Dailey, C. Joyce, M.J. Lyons, M. Kamachi, H. Ishi, J. Gyoba, and G.W. Cottrell, "Evidence and a computational explanation of cultural differences in facial expression recognition," Emotion, vol.10, no.6, pp.874–893, Dec. 2010.
[12] F.A.M. da Silva and H. Pedrini, "Effects of cultural characteristics on building an emotion classifier through facial expression analysis," Journal of Electronic Imaging, vol.24, no.2, p.023015, March 2015.
[13] G. Ali, M.A. Iqbal, and T.-S. Choi, "Boosted NNE collections for multicultural facial expression recognition," Pattern Recognition, vol.55, pp.14–27, July 2016.
[14] T. Kanade, J.F. Cohn, and T. Yingli, "Comprehensive database for facial expression analysis," Proc. 4th IEEE Int. Conf. Automatic Face and Gesture Recog., Grenoble, France, pp.46–53, March 2000.
[15] N. Aifanti, C. Papachristou, and A. Delopoulos, "The MUG facial expression database," Proc. 11th Int. Work. Image Analysis for Multimedia Interactive Services, Desenzano, Italy, pp.1–4, April 2010.
[16] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," Proc. 3th IEEE Int. Conf. Automatic Face and Gesture Recog., Nara, Japan, pp.200–205, March 1998.
[17] M. Biehl, D. Matsumoto, P. Ekman, V. Hearn, K. Heider, T. Kudoh, and V. Ton, "Matsumoto and Ekman's Japanese and Caucasian Facial Expressions of Emotion (JACFEE): Reliability Data and Cross-National Differences," Journal of Nonverbal Behavior, vol.21, no.1, pp.3–21, March 1997.
[18] L.F. Chen and Y.S. Yen, Taiwanese facial expression image database, Brain Mapping Laboratory, Institute of Brain Science, National Yang-Ming University, Taipei, Taiwan, 2007.
[19] A. Fierro-Radilla, K. Perez-Daniel, M. Nakano-Miyatakea, H. Perez-Meana, and J. Benois-Pineau, "An effective visual descriptor based on color and shape features for image retrieval," Proc. 13th Mexican Int. Conf. Artificial Intelligence, Tuxtla, Mexico, vol.8856, pp.336–348, Nov. 2014.
[20] S. Fu, H. He, and Z.-G. Hou, "Learning race from face: A survey," IEEE Trans. pattern analysis and machine intelligence, vol.36, no.12, pp.2483–2509, May 2014.
[21] P. Kakumanu, S. Makrogiannis, and N. Bourbakis, "A survey of skin-color modeling and detection methods," Pattern Recognition, vol.40, no.3, pp.1106–1122, March 2007.

Gibran Benitez-Garcia received his B.S. and M.S. degrees from the Mechanical Engineering School of the National Polytechnic Institute, Mexico City in 2011 and 2014, respectively. He received his Ph.D. degree from the University of Electro-Communications, Japan in 2017. He is currently a postdoctoral fellow in the Intelligent Information Media Laboratory at Toyota Technological Institute in Nagoya, Japan. His research interests include face and facial expression recognition, automatic human-behavior detection, pattern recognition and computer vision.

Tomoaki Nakamura received his B.E., M.E., and Ph.D. degrees from the University of Electro-Communications in 2007, 2009, and 2011, respectively. From 2011 to 2012, he was a research fellow of the Japan Society for the Promotion of Science. In 2013, he worked for Honda Research Institute Japan Co., Ltd. He is currently an assistant professor at the University of Electro-Communications. His research interests are intelligent robotics and machine learning. He is a member of the Robotics Society of Japan and the Japanese Society for Artificial Intelligence.

Masahide Kaneko received his B.E., M.E., and D.E. degrees from the University of Tokyo, Japan in 1976, 1978, and 1981, respectively. From April 1981 to March 1994, he was with the Research and Development Laboratories of Kokusai Denshin Denwa Co., Ltd. (KDD). From April 1994 to March 1997, he was an associate professor of Department of Information and Communication Engineering, the University of Tokyo. In April 1997, he was reinstated in the Research and Development Laboratories of KDD. In April 1998, he joined the University of Electro-Communications as an associate professor. He is currently a professor of Department of Mechanical Engineering and Intelligent Systems. His research interests include the image coding, 3D image processing, processing of facial image information, and active interaction between humans and intelligent robots. He is a fellow of the Institute of Image Information and Television Engineers and a member of IEEE, the Information Processing Society of Japan, the Robotics Society of Japan, and Japan Academy of Facial Studies.