A facial expression image database and norm for Asian population: A preliminary report

Chien-Chung Chen*[a], Shu-ling Cho[b], Katarzyna Horszowska[c], Mei-Yen Chen[a]; Chia-Ching Wu[a], Hsueh-Chih Chen[d], Yi-Yu Yeh[a], Chao-Min Cheng[a]

[a] Department of Psychology, National Taiwan University
[b] Department of Clinical Psychology, Fu-Jen Catholic University
[c] Department of Linguistics, National Taiwan University
[d] Department of Educational Psychology, National Taiwan Normal University

ABSTRACT

We collected 6604 images of 30 models in eight types of facial expression: happiness, anger, sadness, disgust, fear, surprise, contempt and neutral. Among them, 406 most representative images from 12 models were rated by more than 200 human raters for perceived emotion category and intensity. Such large number of emotion categories, models and raters is sufficient for most serious expression recognition research both in psychology and in computer science. All the models and raters are of Asian background. Hence, this database can also be used when the culture background is a concern. In addition, 43 landmarks each of the 291 rated frontal view images were identified and recorded. This information should facilitate feature based research of facial expression. Overall, the diversity in images and richness in information should make our database and norm useful for a wide range of research.

Keywords: emotion, facial features, image processing, face perception

1. INTRODUCTION

A facial expression refers to a combination of the movements and states of facial muscles. Facial expression is perhaps the most efficient accurate indication of not only emotion, such as happiness, anger, or sadness[1-4], but also cognitive status such as concentration or boredom. An ability of correctly interpreting facial expressions allows coordinative interactions among individuals as they respond to a challenge or an opportunity in a social environment. Such ability is considered as an important social skill in many cultures. In the past few years, much progress has been made on automatic facial expression detection and recognition. Such progress allows a computer based expression recognition systems that can identify the facial expression of a person with sufficient accuracy. Hence, automatic facial expression recognition systems have been applied in more and more areas, such as computer assisted patient care, automatic tutoring, law enforcement, interactive entertainment, subjective preference studies, and consumer behavior.

Current major automatic expression recognition methods analyze action units in a facial image to determine the facial expression on that face[5-9]. An action unit is an activity of a facial muscle or muscles that causes an observable movement of some portion of the face[10]. Ekman & Friesen[11] identified 44 different action units. The facial expression of each basic emotion can be described by the movement of a set of action units. Since the action units provide an objective and comprehensive way for describing facial expressions, it is not surprising that analyzing action units in an image is always a critical component of current major automatic expression recognition systems.

As the automatic expression recognition becomes popular, a limitation of the action unit based system also becomes apparent. While it is argued that the action units used for each expression may be universal[12], even the most enthusiastic advocator of universalism of facial expression would agree that the intensity of expression, or the magnitude of movement in action units, may vary from culture to culture[2,13]. Such culture difference occurs not only in how to show an expression but also in how to judge an expression. One expression that is regarded to be a clear indication of a certain

* c3chen@ntu.edu.tw; phone 886 2 33663099; fax 886 2 23639909; http://vnl.psy.ntu.edu.tw.

Image Quality and System Performance VI, edited by Susan P. Farnand, Frans Gaykema,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 7242, 72421D · © 2009 SPIE-IS&T
CCC code: 0277-786X/09/$18 · doi: 10.1117/12.806130

emotion in one culture may be considered as quite ambiguous in another culture. Hence, observers from different cultures may have a great discrepancy in classifying facial expressions. For instance, some images in the widely used POFA database[14] that were classified as "anger" consistently by North Americans could be considered as a representation of a number of negative emotions, such as sadness, disgust, or contempt, but just not anger by Taiwanese observers[15]. Due to such culture difference, an algorithm builds on an image database and norms developed in one culture may not perform ideally when used in another culture. Hence, before applying an automatic expression recognition algorithm to a new population, it may be necessary to test the algorithm with the image database and norms created in the local environment.

Many popular publicly available facial expression databases may not be appropriate for the testing purpose for Asian population. One reason is that those databases contained few Asian models. For instance, The JACFEE database[13] had only three Asian models. The popular Kanade-Cohn[16] had only three "Asian or Hispanic" models among more than 200 total models in that database. The MMI database[17] contains six Asian models. Notice that neither the Kanade-Cohn[16] nor the MMI[17] database contains the human rating of the images in the database. Hence, it is unknown how intense the facial expression in an image is. It is even not clear whether human observers would perceive the same emotion in an image as it is labeled in the database. Thus, it is difficult to make a comparison between machine and human performance with these databases. As a result, missing human rating information makes these databases unsuitable for any study where the human factor is a concern.

The Japanese Female Facial Expression[18] (JAFFE) is perhaps most comprehensive facial expression database for Asian faces. It contains 10 Japanese female models in seven facial expressions. Each image in JAFFE was rated by 50 observers for valence in each basic emotion category. However, all the models in JAFFE were female. Such gender bias may limit its application.

Our purpose of here is not only to establish an image database of facial expression with Asian models but also a norm of emotion judgment on the images in the database by Asian observers. Such database should provide a suitable tool not only for developing an automatic expression recognition algorithm whose performance matches that of human observers but also for facial expression and nonverbal emotion studies with human observers in general.

2. METHODS

2.1 Image acquisition

Thirty models were recruited from either professional theatric groups or theatric schools. There were 14 males and 16 females with age ranged from early 20s to 60s in our models. Before each image acquisition session, the model was informed about the purpose of the project and gave consent for non-profit use of his or her own images. The models were asked to remove their glasses or other items that might block part of their faces.

The models were asked to perform facial expressions for nine emotions: happiness, anger, sadness, disgust, fear, surprise, contempt, satisfaction and neutral. The first six are the basic emotions proposed by Ekman, Friesen & Ellsworth[19]. Contempt and Satisfaction were in the extended list of distinct emotions proposed by Ekman[20]. Contempt was also considered as one of the basic emotions by Izard[21]. There is evidence that contempt and disgust may be mediated by different neural substrates[22]. Satisfaction is a low arousal positive emotion. It is suggested that satisfaction and happiness are two distinct emotion types in Asian population[23]. However, as discussed later in the paper, our result did not support this distinction. Thus we merged the images for satisfaction into the happiness category.

For the six basic emotions on Ekman's list[19], the models were asked to study the facial action coding system[11] (FACS) with the help of an FACS coder and then to perform accordingly. In addition to the FACS based performance, the models were also given a scenario for each emotion category and asked to perform the expression that is appropriate for the emotion in that scenario. After each performing, the models were asked to write down the type and the intensity of the expression they just performed.

The camera and the video camera were placed 1.5 meters away from the models. The models were informed before each taking. The still image camera, which had a better spatial resolution, took pictures continuously at about 1 second interval. The resolution of the still images was 2560(H) x 2048(V). Totally, there were 6604 still images acquired. The video camera streamed at 30 frames per second.

2.2 Human evaluation

There were two stages of rating. The first stage of rating was applied to all 6604 acquired still images. Ten paid raters received training in the FACS coding system[10,11] until they can achieve 80% accuracy in identifying an expression in the POFA database before they started rating images in our database. Each of the 10 raters then rated 1320 to 1325 of all 6604 acquired images. Each image was rated by at least two raters. Figure 1 shows the interface we used for rating. The test image was shown on the left side of the display. A list of possible emotion categories were displayed on the right side of the display. A slider bar was placed adjacent to each emotion name. The task of the raters was to decide the category and the intensity of the emotion that a test image represented by moving the slider in a slider bar to the desired position. The slider on the left end of the slider bar denoted zero intensity while the slider on the left end denoted the maximum intensity of the corresponding emotion felt by the evaluator and was assigned an intensity value 10. The slider placed between the two extremes was assigned a value proportional to its position relative to the two extremes. The raters were asked to give a non-zero intensity in only one emotion category for each image. After each image was rated by at least two raters, we then selected images for the second stage of rating. First, the images were screened for consistency. That is, the images should be classified into the same emotion category by all the raters for that image and the range of the intensity rating across raters should be less than five in the 10-point scale we used. We then selected 12 models whose images passed the consistency check in at least six of seven possible emotion categories. From these selected models, 460 images with the smallest range of rated intensity were selected for the second step of rating.

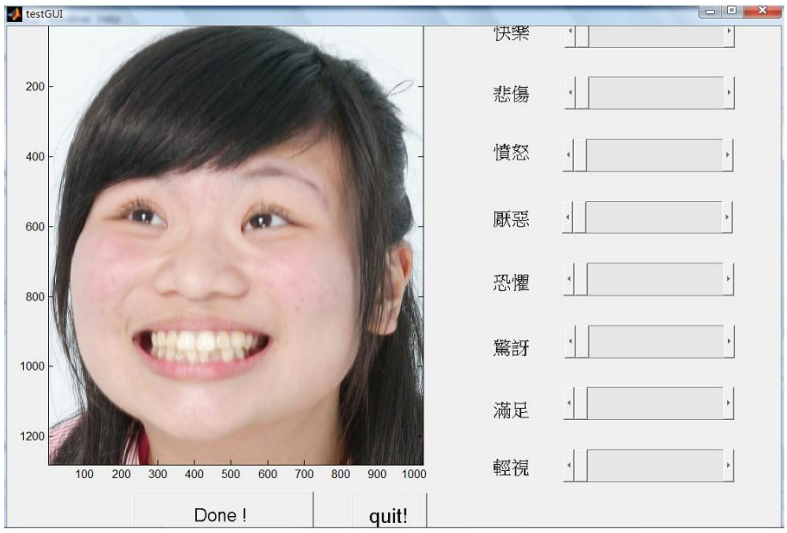

Figure 1. Interface for image rating.

Two hundred and eight raters (147 females, 61 males; mean age: 20.43 yr, standard deviation: 3.16 yr; range: 18~28 yr) from three universities (National Taiwan university, Fu-Jen Catholic University and National Taiwan Normal University) in northern Taiwan participated in the second stage of rating. The raters were randomly assigned in groups of ten for each evaluation session. In each session, the raters first gave informed consent and received instruction for the rating procedure. The interface for the rating was the same as that for the first stage. However, different from the first stage, the raters in the second stage was allowed to give a non-zero intensity in more than one emotion category.

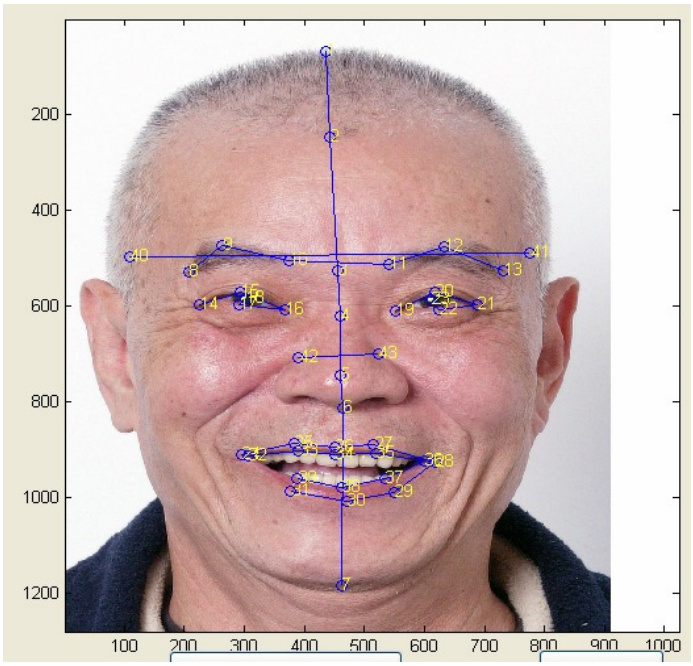
Figure 2. The landmarks used for feature identification. Each circle denotes one landmark point. The connecting lines are shown here to help users to group and comprehend landmarks.

2.3 Feature marking

A facial expression is a combination of the movements and states of facial muscles that

indicates certain type of emotion and cognitive state. Different expressions involve different muscle movements and in turn result in different change in face features. Hence, the image analysis of face features, such as action units, is essential for facial expression recognition[5-11].

To facilitate feature analysis, we had a trained technician to mark the position of 43 landmarks on the faces of the acquired images. At this time, we have marked 291 frontal view images from the 460 images that received the second stage rating. Figure 2 shows the marked points. These 43 marks are organized in groups to aid human users to distinguish and identify them. The first group (Point 1-7) starts from the top of the head, passed through the hair line, the brow, the nasal bridge, to the chin. The second group (Point 8-13) contains points that bisect eyebrows above both eyes. The third group (Point 14-23) includes the pupils, the left and the right corners, and the middle points of the upper and lower eyelids of both eyes. The fourth groups (Point 24-31) contains points that at the left and the right corners and the center of the outer lip edge and those that are half way between the center and the corners at both the upper and lower outer lip edges while the fifth group (Point 32-39) contains the corresponding points at the inner lip edge. The sixth group (Point 40-43) measures the width of the forehead and the nasal bridge. The coordinates of these marks were recorded and is available in the database. In addition to these landmarks themselves, it is easy to calculate the distance between the landmarks. Since a muscle movement often changes the relative positions of neighboring landmarks, the distance measurement among landmarks should be able to catch the change of action units that are not covered by the landmarks themselves. For instance, the change of the distance between the points at the lower eyelid and at the corner of the outer lip can signal the movement of the cheek muscles and the corresponding actions units.

3. HUMAN EVALUATION RESULTS

3.1 Data Analysis

There were two approaches to assign an emotion category to an image. The first one was based on averaged intensity score in each emotion category. In this approach, for each image, we averaged the intensity score in each emotion category across raters. We then assigned an emotion category to an image if the image had the greatest averaged intensity score in that category. For the convenience of discussion, we called the emotion category assigned with this approach as the intensity category of the image. The second approach was based on the categorization performance of individual raters. In this performance, we first determined the emotion categorization at each individual level. An image i belonged to an emotion category j for a particular rater if she gave that emotion category the greatest intensity score. We then computed the number of the raters that assigned each of the emotion categories to the image. The frequency was then scaled by the number of raters to get the probability of assigning emotion category j to the image i, p_{ij}. We then determined the emotion category at the group level of the i-th image as the one with the greatest probability p_{ij}. This approach determined the emotion category by a plurality vote. Hence, we will call the emotion category of an image determined with this approach as the plurality category. The plurality category is less affected by outliers. Hence, if not mentioned otherwise, the term "category" in the following discussion refers to the plurality category.

In additional to the categorization responses, the probability of assigning a category j to an image i can be used to evaluate the consistency among observers in emotion judgment. The entropy level of image i was calculated as

$$\Sigma_j[-p_{ij}*\log_2(p_{ij})] \quad (1)$$

where j denoted emotion categories. The entropy is zero if all raters agreed in their emotion categorization. A large entropy level means a great amount of inconsistency among the raters.

3.2 Categorization responses

For the great majority of images, the plurality category agreed with the intensity category: only 13 out of 460 images (2.8%) showed discrepancy between the two designations. Such discrepancy occurred when the expression in that image had a low intensity. As a result, an observer may feel the expression ambiguous and thus may give that image intensity scores equally distributed in two or more categories. Hence, random variability among observers would dictate the result of categorizations. For instance, an observer is very likely to assign a low intensity score to a random category to a neutral expression. As a result, the plurality category for a neutral face would be completely random. Indeed, all the thirteen images showing discrepancy in two categorizations has a low intensity score for the plurality category (ranged

Table 1. The frequency of images in each plurality category breaks down by models.

Model No.	Gender	Happiness	Sadness	Anger	Disgust	Fear	Surprise	Contempt
1	F	3	3	1	3	0	3	0
2	M	11	4	0	1	0	1	0
3	M	3	1	7	2	2	5	2
4	M	6	4	5	4	1	3	2
5	M	6	1	3	3	0	5	0
6	M	16	2	0	1	1	11	3
7	F	51	17	13	2	0	5	0
8	F	36	3	1	2	0	4	3
9	M	26	1	2	1	0	19	1
10	F	35	3	8	8	10	6	2
11	M	8	2	1	6	1	5	9
12	F	15	2	4	2	1	5	9
	sum	**216**	**43**	**45**	**35**	**16**	**72**	**31**

1.18-2.24, mean=2.09 on a 10-point scale; or in 0.04 to 22 percentile) and high intensity entropy (ranged 1.06-2.82, mean 1.8; or in 46-99 percentile).

The human observers respond little to the category of Satisfaction. Only two out of 460 images was assigned to satisfaction. In addition, these two images were considered as happiness by a large proportion of raters. Hence, while it is suggested that Asians may react differently in the situation where they feel satisfied from the situation where they feel happy in general[23], our raters did not perceive this difference. Hence, we decided to merge these two categories. The two images in question were then reassigned to the Happiness category in the database.

The category Contempt was showed a large variability among our raters in image assignment. Notice that, in the literature, whether contempt should be considered as one of the basic emotions is still in debate. From the performance of our raters, we found a high correlation in intensity score between Contempt and Disgust ($r=0.38$, $p<0.0001$). In addition, for all images assigned to Contempt, the Disgust was always the second best choice. However, the opposite trend was true only for 37% of images in Disgust. Hence, it is likely that, for Taiwanese people, contempt is only a subset of disgust.

Table 1 shows the frequency of images in each plurality category breaks down by models. Five models had images in all seven emotion categories. The other three models lacked image in one category: two lacked Fear while another one lacked Anger. The rest four models lacked images in Fear and Contempt. Nevertheless, all models had images in at least five basic emotion categories. Such coverage is in par with some major facial expression databases[13,14].

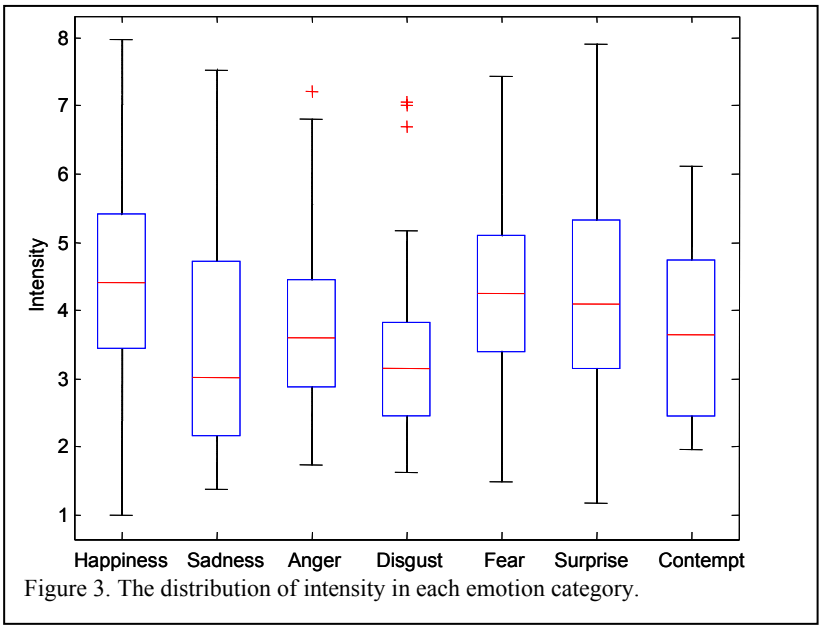

Figure 3. The distribution of intensity in each emotion category.

3.3 Intensity responses

The box plots in Figure 3 summarizes the distribution of the intensity score in each category. The lower and upper edge of each box of the box plot denotes the intensity score of 25% and 75% percentile respectively and the horizontal line in the box denotes the median. The bars below and above a box denote the range of the

intensity score in that category excluding the outliers determined with the Gaussian assumption of the data. These apparent outliers were denoted by crosses.

The averaged intensity score for all images had a mean 4.11 and a standard deviation 1.49. The median was 3.98 with the 25% and 75% percentile at 2.91 and 5.18 respectively. Such a diversity of intensity score should satisfy a wide range of research needs. The range of intensity score was similar for all categories except Disgust. The Disgust distribution concentrated more on the lower end of overall intensity range. However, there were still images with high Disgust intensity scores in the database as indicated by the red crosses in Figure 3.

3.4 Entropy responses

We used entropy as a metrics of individual difference in category assignment. Zero entropy means that all observers assigned the same category to this image. After merging Satisfactory with Happiness, there were seven possible categories for each image. The entropy may reach its upper limit if the raters evenly assigned all seven categories to an image. That is, the largest possible entropy, calculated with Eq. 1, is $7*(-1/7*\log_2(1/7)) = 2.81$. Of course, due to random variability, it is quite unlikely to get entropy at this upper limit. Instead, a completely random category assignment will yield a distribution of entropy. We estimated this null distribution with a Mote Carlo simulation of 10,000 raters. For a completely random assignment, there was 95% chance that the entropy was above 2.76. Such high entropy can occur in real life. For instance, assume that we asked a rater to assign a category to an image of neutral expression (or an expression of very low intensity). There was no a priori reason for a rater to assign a particular category to this image. Hence, when we forced raters to make a choice, the raters may just make the assignment randomly. In this case, the entropy will be close to that for completely random assignment. Figure 4 shows the entropy distribution of our data. The mean entropy was 0.99 with a standard deviation 0.68 while the median was 1.02. The 25% and 75% percentile entropy was 0.35 and 1.54 respectively. The median was close to 1. This suggests that most images have only one dominate emotion category. The greatest entropy was 2.44, smaller than 2.76. Hence, none of the images rated received random assignment.

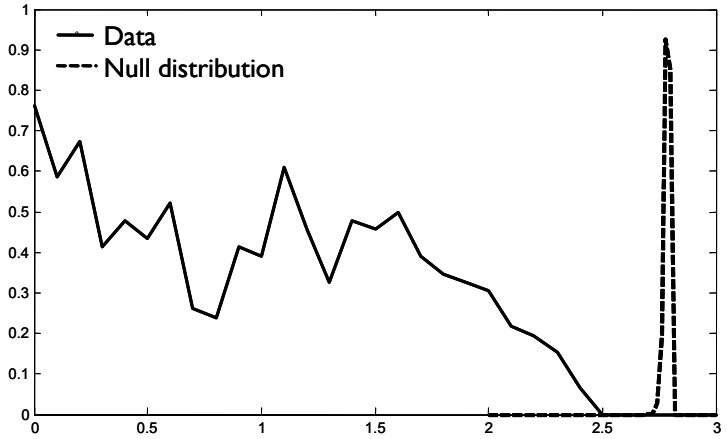

Figure 4. The distribution of entropy. Solid curve: the empirical distribution from the data. Dotted curve: the null distribution under the null hypothesis that the raters rated images randomly. The probability density shown in this figure was normalized for visualization.

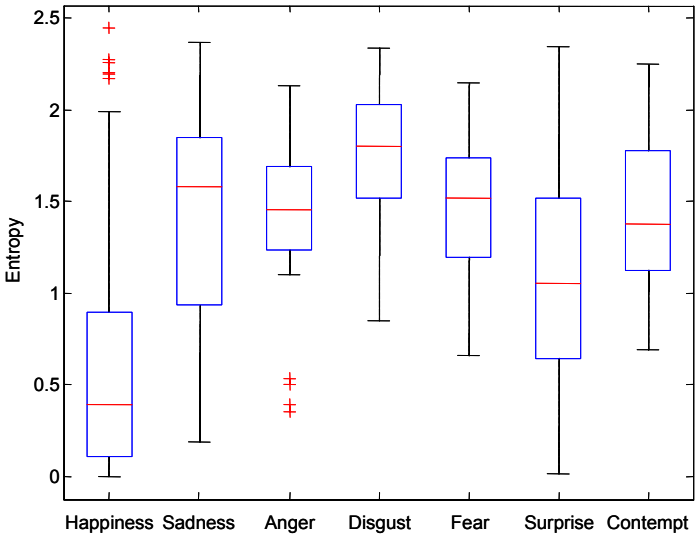

Figure 5. The distribution of entropy in each emotion category.

Figure 5 shows the distribution of entropy in each category in box plots

similar to that of Figure 3. The mean entropy was between 0.56 (Happiness) and 1.74 (Disgust). Happiness, Sadness and Surprise had a wide spread of entropy values. Anger has an entropy distribution similar to that of Disgust and Contempt. However, there are also images with low entropy (Red crosses) in these categories.

4. DATABASE

There are 6604 images of 30 models in eight emotion categories (Satisfaction was merged with Happiness). Among them, 406 images received human evaluation and will be the major components of the database and 291 images received feature marking. Figure 6 shows an example of images of one model. For the convenience of users, we provided a searchable interface written in PHP with an Apache webserver. Thus, it is possible to access the database, whose back-end software for data storage and retrieval was written in MySQL, through a web browser. This web-based user interface provided functions such as searching facial expression images with a given set of parameters.

There are four types of information for each image in our database: (1) Image acquisition, (2) viewpoint, (3) human rating, and (4) feature marking. Image acquisition information is available for all 6604 images. This information group includes the image itself, the identity number, gender, and age of the model, and the type of expression intended by the model. The viewpoint and the human rating groups are available to the 406 images received human rating. The former includes the view (frontal, 3/4, or profile), and, if there is any, the pitch and the roll of the head movement. The human rating information includes both categorical and intensity rating data. The categorical rating information includes (1) the plurality category; and (2) the intensity category. The intensity rating includes (1) the average intensity of the image in both plurality and the intensity category; (2) the standard deviation of the average intensity in both category assignment; (4) the entropy that the individual difference in category assignment; and (5) the averaged intensity in each emotion category. Finally, the feature marking information, available to 291 images, contains the horizontal and the vertical coordinates of each of the 43 landmarks on a frontal view face.

					
Category	happiness	sadness	anger	Disgust	Surprise
Intensity	4.63	2.08	4.46	1.60	5.56
Entropy	0.80	1.65	1.31	2.29	0.76
					
Category	happiness	sadness	anger	Disgust	Surprise
Intensity	3.78	4.41	3.66	2.30	2.34
Entropy	1.02	0.93	1.46	2.24	1.88

Figure 6. Example images from one model with different expression and view point.

5. CONCLUSION

In this project, we collected 6604 images of 30 models in eight types of facial expression: happiness, anger, sadness, disgust, fear, surprise, contempt and neutral. Among them, 406 most representative images from 12 models were rated by more than 200 human raters for perceived emotion category and intensity. The number of emotion types, models and raters should be large enough for serious expression recognition research both in psychology and in computer science. All the models and raters are of Asian background. Hence, this database can also be used when the culture background is a concern. In addition, 43 landmarks each of the 291 rated frontal view images were identified and recorded. This information should facilitate feature based research of facial expression. Overall, the diversity in images and richness in information should make our database and norm useful for a wide range of research.

AKNOWLEGEMENT

This study is supported by National Science Council (Taiwan), NSC 96-2752-H-002-004-PAE to CCM, NSC 96-2752-H-002-007-PAE & NSC 97-2410-H-002-158-MY2 to CCC.

REFERENCES

[1] Ekman, P., "Expression and the nature of emotion," *Approach to Emotion*, Scherer, K. & Ekman, P. (Eds.), 319-344, Lawrence Erlbaum, Hillsdale, New Jersey, USA, (1984).

[2] Ekman, P., "Facial expression & Emotion," American Psychologist **48**, 384-392 (1993).

[3] Izard, C. E., [The face of emotion], Appleton, New York, USA, (1971).

[4] Haidt, J. and Keltner, D., "Culture and facial expression: Open ended methods find more faces and a gradient of universality," Cognition and emotion **13**, 225-266 (1999).

[5] Bartlett, M. S., Hager, J. C., Ekman, P. and Sejnowski, T. J., "Measuring facial expressions by computer image analysis," Psychophysiology **36**(2), 253-263 (1999).

[6] Bartlett, M. S., Littlewort, G., Frank, M. G., Lainscsek, C., Fasel, I. and Movellan, J., "Fully automatic facial action recognition in spontaneous behavior," Proc. IEEE Conf. Automatic Face & Gesture Recognition, 223-230 (2006).

[7] Tian, Y. L., Kanade, T. and Cohn, J. F., "Facial Expression Analysis," *Handbook of Face Recognition*, Li, S.Z. & Jain, A. K., (Eds.), 247-276, Springer, New York, USA, (2005).

[8] Pantic, M. and Rothkrantz, L. J. M., "Facial action recognition for facial expression analysis from static face images," IEEE Trans. on Systems, Man and Cybernetics - Part B **34**(3), 1449-1461 (2004).

[9] Cohn, J. F. and Ekman, P., "Measuring facial actions," *The New Handbook of Methods in Nonverbal Behavior Research*, Harrigan, J.A., Rosenthal, R. & Scherer, K., (Eds.), 9-64, Oxford University Press, New York, USA, (2005).

[10] Ekman, P., Friesen, W. V. and Hager, J. C., [The Facial Action Coding System (2nd. Ed.)], Research Nexus eBook, Salt Lake City, USA, (2002).

[11] Ekman, P. and Friesen, W. V., [Facial action coding system: A technique for the measurement of facial movement], Consulting Psychologists Press, Palo Alto, CA, (1978).

[12] Ekman, P., "Universals and cultural differences in facial expressions of emotion," *Nebraska Symposium on Motivation*, J. Cole (Ed.), 19, 207-283, University of Nebraska Press, Lincoln, NE, (1972).

[13] Matsumoto, D. and Ekman, P., [Japanese and Caucasian facial expressions of emotion (JACFEE)] (Slides), Intercultural and Emotion Research Laboratory, Department of Psychology, San Francisco State University, San Francisco, CA, (1988).

[14] Ekman, P. and Friesen, W. V., "Pictures of facial affect," Human Interaction Laboratory, University of California Medical Center, San Francisco, (1976).

[15] Cho, S. L, Yang, T. M. and Chen, H. C., "Facial expression perception in schizophrenic patients," The College of Medicine Fu-Jen University Technical Report, (2005).

[16] Kanade, T., Cohn, J. F. and Tian, Y., "Comprehensive database for facial expression analysis," Proc. IEEE Conf. Automatic Face & Gesture Recognition, 46-53 (2000).
[17] Pantic, M., Valstar, M. F., Rademaker, R. and Maat, L., "Web-based database for facial expression analysis," Proc. IEEE Int'l Conf. Multimedia and Expo, 317-321 (2005).
[18] Lyons, M. J., Akamatsu, S., Kamachi, M. and Gyoba, J., "Coding Facial Expressions with Gabor Wavelets," Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, 200-205 (1998).
[19] Ekman, P., Friesen, W. V. and Ellsworth, P., "What emotion categories or dimensions can observers judge from facial behavior?" *Emotion in the human face*, P. Ekman (Ed.), 39-55, Cambridge University Press, New York, USA, (1982).
[20] Ekman, P., "Basic Emotions," *The Handbook of Cognition and Emotion*, T. Dalgleish and T. Power (Eds.), 45-60, John Wiley & Sons, Ltd, Sussex, U.K., (1999).
[21] Izard, C. E., [Human emotions], Plenum Press, New York, USA, (1977).
[22] Sambataro, F., Dimalta, S., Di Giorgio, A., Taurisano, P., Blasi, G., Scarabino, T., Giannatempo, G., Nardini, M. and Bertolino, A., "Preferential responses in amygdala and insula during presentation of facial contempt and disgust," European Journal of Neuroscience **24**(8), 2355-62 (2006).
[23] Tsai, J.L., Knutson, B. and Fung, H.H., "Cultural variation in affect valuation," Journal of Personality and Social Psychology **90**, 288-307 (2006).