

# AN INFANT FACIAL EXPRESSION RECOGNITION SYSTEM

Chiung-Yao Fang(方瓊瑤), Liu-Jia Huang(黃律嘉), and Sei-Wang Chen(陳世旺)

<sup>1</sup> Dept. of Computer Science and Information Engineering,  
National Taiwan Normal University, Taipei  
E-mail: violet@csie.ntnu.edu.tw

## ABSTRACT

This paper presents a vision-based infant facial expression recognition system, which can be applied to infant monitoring systems to reduce the take-care load of the caregivers. In this study, a video camera positioned above the infant's crib captures video. The proposed system consists of two stages: infant face detection and infant facial expression classification. Here we apply an improved Locus model to detect the infant skin color for infant face detection. The proposed face detection method can be applied to different infants or various lighting environments. Moreover, a principal component analysis method is used to extract the features for the infant facial classification. In this study, we classify the infant facial expression into seven classes, including sleeping, dazing, crying, laughing, yawning, sneezing and vomiting. The experimental results show that the proposed method is robust and efficient.

**Keywords:** *Infant Face Detection; Facial Expression Recognition; Locus model; Principal Component Analysis*

## 1. INTRODUCTION

Infants are susceptible to illness and to injuries sustained due to the negligence of the babysitter. Therefore some infant care products for home safety have been designed or developed recently, such as infant movement sensor and infant radio/video monitor. An infant movement sensor consists of an under-the-mattress sensor pad which can detect the infant's slightest movements, and will transmit a no-movement alarm to the caregivers.

An infant radio/video monitor is a unidirectional video/radio transmitter/receiver system used for remotely watching/listening. The transmitter that comes with a camera or a microphone is placed near the infant, while the receiver equipped with a monitor or a speaker is kept near the caregiver. Some of the infant monitors are bi-directional, which actually allow for two-way communication.

We believe that to develop an automatically vision-based infant monitoring system is an intelligent and

efficient method to reduce the take-care load of the caregivers. Compare with the above infant radio/video monitor, the automatically vision-based infant monitoring system can actively detect the infant accidents and send the warning messages to the caregivers.

The infant accidents may include infants cry or vomit, infant faces or bodies are occluded by foreign bodies, or infants fall down the cribs. Thus the infant face detection and infant facial expression recognition system plays an important role in the vision-based infant monitoring system. If we set a camera on the crib or before the baby carriage, then once the infant is crying or vomiting, the system can output message to warn the caregivers to avoid the infant suffocation. Here in this study the term 'infants' indicates the infant whose ages are between one to six months.

Many facial expression recognition methods have been proposed recently. Silva et al. [1] proposed a cloud basis function (CBF) neural network to recognize six universal facial expressions from static images. These six facial expressions include fear, surprise, sad, disgust, and happy. Xie and Lam [2] used a spatially maximum occurrence model (SMOM) to represent facial expressions, and then used an elastic shape-texture matching (ESTM) algorithm to measure the similarity between frames based on the shape and texture information. The similarity can be used to recognize the facial expressions. Gu et al. [3] proposed a hybrid facial expression recognition framework in the form of a novel fusion of statistical techniques and the known model of a human visual system.

However, most of them focus on recognizing facial expressions of adults. Compared to adults, the head poses of the infants are more difficult to locate because we cannot ask the infant do not move their heads. In fact, very few infant facial expression recognition methods have been proposed to date.

Pal et al. [4] used the position of the eyebrows, eyes, and mouth to estimate the individual motions in order to classify infant facial expressions. The various classes of facial expressions include anger, pain, sadness, hunger, and fear. The features they used are the local ones.

However, the facial expressions are relative to the change of not only the facial features described above but also the lighting shadows of the facial skin. Here we proposed some global features which are more suitable to recognize the infant facial expressions. This study classifies the infant facial expression into seven classes, including sleeping, dazing, crying, laughing, yawning, sneezing and vomiting. Now, we introduce the system flowchart of the proposed infant facial expression recognition systems.

## 2. SYSTEM FLOWCHART

As mentioned above, the video camera is positioned above the crib or baby carriage to capture infant faces. Figure 1 shows some input examples and Figure 2 shows the flowchart of the infant facial expression recognition system. The system is divided into two stages, one is the database construction stage; the other is the infant facial expression recognition stage.



Fig. 1: Input examples.

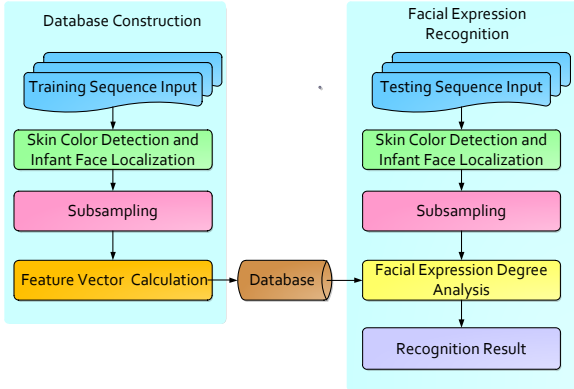


Fig. 2: Flowchart of the facial expression recognition system.

In the database construction stage, once the training sequences are input into the system, the skin color areas of the input frames are detected and the infant faces are then located. Here we use an improved Locus model to detect the skin color areas and use the connected component technique to locate the infant face. For improving the processing efficiency, the infant face region is sub-sampled to a  $40 \times 30$ -pixel size, and then a principal component analysis (PCA) method is applied to extract the features of the infant facial expressions. These feature vectors are then stored in a database for the next stage.

In the infant facial expression recognition stage, once the testing sequences are input into the system,

their process is similar to the database construction stage. The skin color areas are detected, the infant face is located and the face region is sub-sampled. Moreover, the facial expression features are also extracted using the same PCA method. These extracted feature vectors are then compared with those stored in the database to select the most similar one. The classification result depends on which class the selected feature vector belongs to.

## 3. SKIN COLOR DETECTION AND INFANT FACE LOCALIZATION

Soriano et al. [5] have proposed a Locus model to detect the skin areas of the frames. Let  $R, G, B$  be the intensity values of a pixel  $p$  in RGB color model, and

$$r = \frac{R}{R+G+B}, \quad g = \frac{G}{R+G+B}.$$

They observe that the range of the skin colors is like an eyebrow area on  $r$ - $g$  plane, except a little circle area the center of which is at  $(0.33, 0.33)$ , shown in Figure 3. The eyebrow area can be bounded by two curves

$$F_1(r) = -1.376r^2 + 1.0743r + 0.1452 \quad (1)$$

$$F_2(r) = -0.776r^2 + 0.5601r + 0.1766 \quad (2)$$

and the circle area can be modeled by

$$F_3(r, g) = (r - 0.33)^2 + (g - 0.33)^2 \quad (3)$$

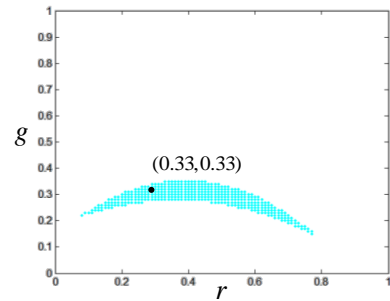


Fig. 3: The Locus model.

Thus the skin color pixel can be defined by

$$S = \begin{cases} 1 & \text{if } ((F_2(r) < g < F_1(r)) \& (F_3(r, g) > T)) \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Here  $T$  is a threshold which indicates the radius of the circle area. If  $S=1$  then  $p$  is a skin color pixel, otherwise it is not. The disadvantage of this Locus model is that some other colors, for example, blue, white, lime green and yellow, will also be extracted. Thus Huang [6] proposed an improved Locus model to solve this problem. He added three criteria as follows:

$$R > G > B \quad (5)$$

$$R - G > 5 \quad (6)$$

$$F_4(r, g) = (r - 0.5)^2 + (g - 0.5)^2 > 0.03 \quad (7)$$

Eqs. 5, 6, and 7 are used to filter the blue, lime green, and yellow pixels, respectively. On the other hand, we found that the value of  $T$  will also affect the results of skin color detection, shown in Figure 4. Here we proposed a method can automatically decide a suitable value of  $T$  for the infant face detection.

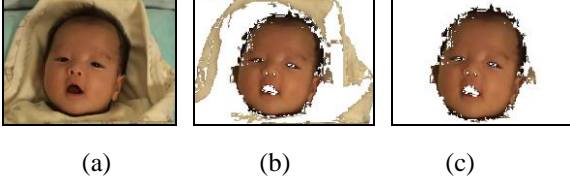


Fig. 4: An example of skin color area detection. (a) the input frame. (b)  $T = 0.0004$ . (c)  $T = 0.009$ .

Since we suppose that input frame contains an infant face, the system first initializes  $T = 0.0004$  and then increases by 0.0001 per iteration until the extracted face region occupied 30% to 40% areas of the whole frame. Figure 5 shows some experimental results of infant face detection. We can observe that these detection results are correct and robust in various situations.

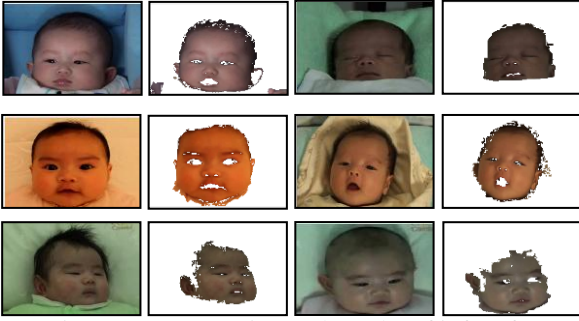


Fig. 5: Some experimental results of infant face detection.

The extracted skin regions will be labeled by a connected component method and the maximum region will be regarded as the infant face region. The system will finally find a minimum bounding rectangle of the face region to extract the infant face. The extracted minimum bounding rectangle of infant face region is then converted to gray scale and reduced its size to  $40 \times 30$  pixels before feature extraction, shown as Figure 6. Figure 6 (a) shows the input frame, and the gray scale image after face detection is shown in Figure 6 (b). Figure 6 (c) shows the result after sub-sampling.

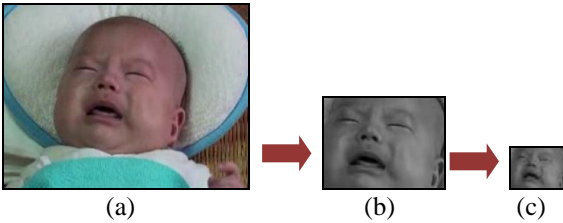


Fig. 6: An example of infant face detection. (a) The input image. (b) The gray scale image after face detection. (c) The result after sub-sampling.

#### 4. FEATURE VECTOR CALCULATION

In this study we use the principal component analysis method (PCA) [7] to extract the facial expression features. Let  $M$  be the number of training frames whose size is  $n$ , and each frame can be represented by an  $n$ -dimension input vector  $\mathbf{x}_k$ , where  $k = 1, \dots, M$ . The PCA method can be used to find an  $n \times m$  transform matrix  $\mathbf{W}$  to transform the  $n$ -dimension input vectors into  $m$ -dimension feature vectors, where  $m \leq n$ .

$$\mathbf{z}_k = \mathbf{W}^T \mathbf{x}_k, k = 1, 2, \dots, M. \quad (8)$$

First, the system calculates the mean of the input vectors,  $\bar{\mathbf{x}}$ ,

$$\bar{\mathbf{x}} = \frac{1}{M} \sum_{k=1}^M \mathbf{x}_k, \quad (9)$$

and the difference vector between the input vector and the mean vector is

$$\boldsymbol{\phi}_k = \mathbf{x}_k - \bar{\mathbf{x}}, k = 1, 2, \dots, M. \quad (10)$$

Then the total scatter matrix of input vectors can be defined as

$$\sum_{T_x} = \sum_{k=1}^M \boldsymbol{\phi}_k \boldsymbol{\phi}_k^T = \sum_{k=1}^M (\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})^T = \mathbf{A} \mathbf{A}^T. \quad (11)$$

where  $\mathbf{A} = [\boldsymbol{\phi}_1 \boldsymbol{\phi}_2 \dots \boldsymbol{\phi}_M]$ . On the other hand, the total scatter matrix of feature vectors is

$$\begin{aligned} \sum_{T_z} &= \sum_{k=1}^M (\mathbf{z}_k - \bar{\mathbf{z}})(\mathbf{z}_k - \bar{\mathbf{z}})^T \\ &= \sum_{k=1}^M (\mathbf{W}^T \mathbf{x}_k - \mathbf{W}^T \bar{\mathbf{x}})(\mathbf{W}^T \mathbf{x}_k - \mathbf{W}^T \bar{\mathbf{x}})^T \end{aligned} \quad (12)$$

From Eqs. (11) and (12), we can obtain that

$$\sum_{T_z} = \mathbf{W}^T \sum_{T_x} \mathbf{W}. \quad (13)$$

If the values of  $\sum_{T_z}$  are large, it means the feature vectors scatter at the feature space widely. Therefore we seek  $\mathbf{W}$  such that  $\sum_{T_z}$  is maximized subject to the constraint that  $\mathbf{W}^T \mathbf{W} = \mathbf{I}$ . This is a Lagrange problem,

$$\max_{\mathbf{W}} \mathbf{W}^T \sum_{T_x} \mathbf{W} - \lambda (\mathbf{W}^T \mathbf{W} - \mathbf{I}). \quad (14)$$

It is known that if  $\lambda_1$  is the largest eigenvalue of  $\sum_{T_x}$ , then its corresponding eigenvector is the first principal component. Similarly, the second principal component is the corresponding eigenvector of the second large eigenvalue  $\lambda_2$ . Using this method, we can obtain  $m$ th principal components to construct a new matrix  $\mathbf{W}'$ , where  $m < n$ .

To decide the number of principal components,  $m$ , is a trade-off. A large  $m$  causes time consuming, and a small  $m$  causes information lost.

Let  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m \geq \dots \geq \lambda_n$  is a sorted sequence of all eigenvalues, in the system  $m$  is automatically decided by the following equation [8].

$$\frac{\lambda_1 + \lambda_2 + \dots + \lambda_m}{\lambda_1 + \lambda_2 + \dots + \lambda_n} \geq 80\%$$

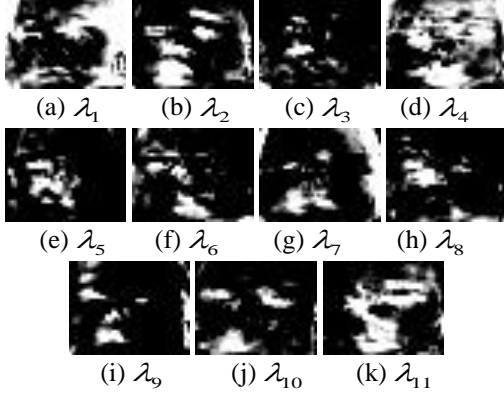


Fig. 7: The examples of eigenfaces corresponding to different eigenvalues, where  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{11}$ .

The eigen vectors of face images are called eigenfaces. Figure 7 shows an example of these eigenfaces. These eigenfaces will be regarded as features stored into the database to represent different facial expressions.

## 5. FACIAL EXPRESSION RECOGNITION

Given a transform matrix  $\mathbf{W}'$ , once a testing frame  $\mathbf{y}$  input into the system, it will be transferred into a  $m$ -dimension feature vector  $\mathbf{v}$  by the following equation.

$$\mathbf{v} = \mathbf{W}'^T \mathbf{y} \quad (15)$$

This feature vector  $\mathbf{v}$  is then compared with the feature vectors stored in the database to find the most similar one, named  $\mathbf{z}^*$ .

$$\mathbf{z}^* = \arg \min_{\mathbf{z}} (d(\mathbf{v}, \mathbf{z}_1), d(\mathbf{v}, \mathbf{z}_2), \dots, d(\mathbf{v}, \mathbf{z}_M)), \quad (16)$$

where  $d$  indicates the Euclidean distance between two vectors. If  $\mathbf{z}^*$  belongs to class  $c$ , then the facial expression class of  $\mathbf{y}$  can be assigned to the same class.

$$c_y = c \text{ if } \mathbf{z}^* \in c, \quad (17)$$

The above classification result is only a temporary result since using only one frame to classify the facial expression is not robust. We believe to integrate the classification results of successive frames should increase the robustness of the system.

Since the input data are successive frames and they may represent the same facial expression during a period of time, we proposed a method which is based on a probability adaptive model to recognize the infant facial expressions. This recognition flowchart is shown in Figure 8.

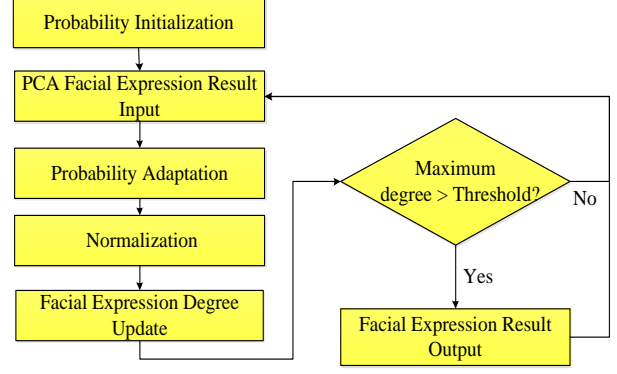


Fig. 8: Flowchart to recognize facial expressions.

Initially, the probabilities of all facial expression classes stored in the database are set to a constant value first. Once a frame input into the system, using its PCA facial expression classification result, the system can adapt the likelihood of each facial expression class. And then the likelihood will be normalized to be the probability of each facial expression class and their corresponding facial expression degrees will be updated. The facial expression degree indicates the accumulative probability of a facial expression until current moment. Once the maximum facial expression degree of one facial expression is greater than a threshold, the system outputs the final facial expression recognition result.

Let the facial expression be  $k$  classes,  $P'(c)$  be the probability of the facial expression class  $c$  at time  $t$ . Initially, the probabilities of all  $k$ th facial expression classes are all set to  $1/k$ . Once a frame  $\mathbf{y}$  of a sequential video at time  $t$  is input into the system and is classified into class  $c_y$  finally, the system updates the likelihood of each class at time  $t+1$  by

$$P^{t+1}(c) = P^t(c) + \frac{\sum_{i=1}^t O_i \times S(i)}{\sum_{i=1}^t S(i)}, \quad (18)$$

where  $O_i = \begin{cases} 1 & \text{if } c_y = c \text{ at time } i \\ 0 & \text{otherwise} \end{cases}$ , and

$$S(t, t') = \frac{a}{\sigma\sqrt{2\pi}} e^{-\frac{(t-t')^2}{2\sigma^2}}.$$

Here  $a$  is a constant and  $S(t, t')$  is a time decay function which calculates the weight of the output at time  $t'$  when the system processes the frame at time  $t$ . We can observe that if  $t'$  is close to  $t$  then the value of time decay function  $S(t, t')$  will be large. In fact, it is a monotonic decreasing function of  $|t-t'|$ .

## 6 EXPERIMENTAL RESULTS

The input data for our system was acquired using the SONY HDR-XR500, which were mounted above a crib or baby carriage. The video was processed on a PC with a 3.1GHz CPU. The input video sequences captured at



30 fps (frames/second) were down-sampled to 6 fps, which is the processing speed of our current system. The frame size is 320 x 240 pixels.



Fig. 9: Some examples of the input video sequences.

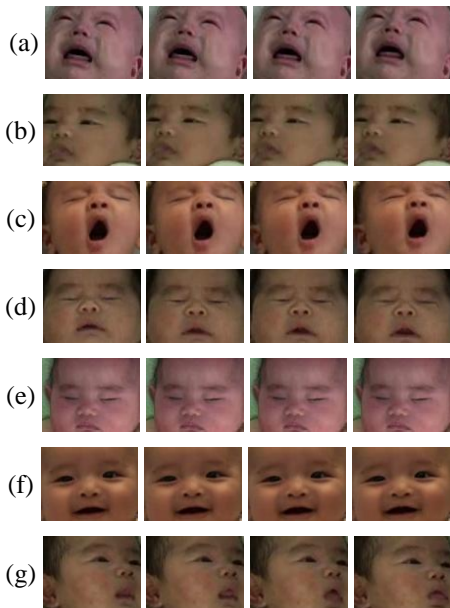


Fig. 10: The corresponding face detection results of the input frames shown in Figure 9.

The first experiment is about the face detection. Figure 9 shows some examples of the input video sequences. These infant facial expression sequences are obtained from different babies, different time, and different background situations, including sleeping, dazing, crying, laughing, yawning, sneezing and vomiting. Moreover, the head of the babies may turn left or right. Figure 10 shows the corresponding face detection results, by minimum bounding rectangles of the face regions, of the input frames shown in Figure 9. We can observe that the face detection results are correct and robust even the infants turn their heads to the left or to the right.

The second experiment shows the infant facial expression recognition results. In this experiment, totally 813 feature vectors are stored in the database. On the other words, totally 813 training frames are used in this experiment, including sleeping (63 frames), dazing (139 frames), crying (168 frames), laughing (142 frames), yawning (118 frames), sneezing (102 frames) and vomiting (81 frames) expressions. These training data include the frames that the infants turn their head to the left or to the right.

Tables 1 to 7 show the accuracy rate of each facial expression respectively. From Tables 1, 3, and 7, we can observe the recognition accuracy rate of crying, yawning, and dazing expressions is all higher than 91%. The vomiting and sneezing expressions (shown in Tables 2 and 4 respectively) are more difficult to recognize since that expression is not obviously distinguishing characteristic. The accuracy rates of sleeping and smiling expressions are more unstable than the other expressions. However, all the recognition accuracy rates are higher than 82%. We think using more training frames will increase the accuracy rate of the facial expression recognition system.

Table 1. The recognition results of crying sequences.




One example of input sequence	No. of testing frames	Correct results	Accuracy rate
	126	121	<b>96%</b>
	200	190	<b>95%</b>
	35	33	<b>94.2%</b>

Table 2. The recognition results of vomiting sequences.



One example of input sequence	No. of testing frames	Correct results	Accuracy rate
	161	135	<b>83.8%</b>
	51	42	<b>82.3%</b>

Table 3. The recognition results of yawning sequences.




One example of input sequence	No. of testing frames	Correct results	Accuracy rate
	130	130	<b>100%</b>
	75	71	<b>94.6%</b>
	21	21	<b>100%</b>

Table 4. The recognition results of sneezing sequences.



One example of input sequence	No. of testing frames	Correct results	Accuracy rate
	226	190	<b>84%</b>
	78	68	<b>87%</b>

Table 5. The recognition results of sleeping sequences.




One example of input sequence	Testing no. of frames	Correct results	Accuracy rate
	155	155	<b>100%</b>
	101	83	<b>82%</b>
	100	100	<b>100%</b>

Table 6. The recognition results of smiling sequences.







One example of input sequence	No. of testing frames	Correct results	Accuracy rate
	159	158	<b>99.3%</b>
	101	83	<b>82%</b>
	71	65	<b>91.5%</b>

Table 7. The recognition results of dazing sequences.

One example of input sequence	No. of testing frames	Correct results	Accuracy rate
	135	130	<b>96.2%</b>
	141	140	<b>99%</b>
	108	99	<b>91.6%</b>

The third experiment shows the infant facial expression recognition results about testing sequence 1. We test a 23-minute sequence, totally 3462 frames. This sequence contains has three facial expressions, including smiling, yawning, and dazing expressions. Some frames

of the sequence are shown in Figure 11. We can observe that the infant's head turn to the left sometimes.

The recognition results are shown in Table 8. The first column contains the three facial expressions of the input sequence. Table 8 shows that only two of the totally 166 frames of the smiling expressions are recognized incorrectly. They are classified to the yawning class. The 229 frames of the yawning expressions are all correctly recognized. However, the frames of dazing expressions will be assigned to different incorrect classes. However, we can observe that the recognition rates of the smiling, yawning, and dazing expressions are still very high. They are 98.79%, 100%, and 98.33%, respectively.

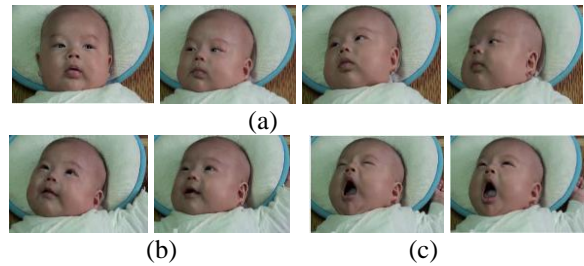


Fig. 11: Some frames of the testing sequence 1. (a) Dazing examples. (b) Smiling examples. (c) Yawning examples.

Table 8. The recognition results of testing sequence 1.

Recognition result \ Actual facial expressions	smiling	yawning	dazing	sneezing	sleeping
smiling	<b>164</b>	<b>2</b>			
yawning		<b>229</b>			
dazing	<b>7</b>	<b>18</b>	<b>2954</b>	<b>24</b>	<b>1</b>

Figure 12 shows some incorrect examples. Figure 12 (a) shows a frame of the smiling expression classified to the yawning class. Figures 12 (b) (c) and (d) are all frames of the dazing expression, but they are classified to the yawning, sneezing, and sleeping expressions respectively.



Fig. 12: Some incorrect examples.

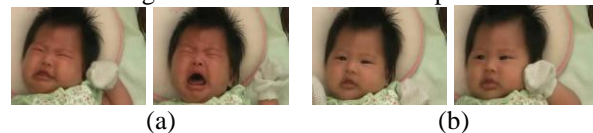


Fig. 13: Some frames of the testing sequence 2. (a) Crying examples. (b) Dazing examples.

The fourth experiment shows the infant facial expression recognition results about testing sequence 2. We test a 22-minute sequence, totally 3345 frames. This sequence contains has two facial expressions, including crying and dazing expressions. Some frames of the

sequence are shown in Figure 13. The recognition results are shown in Table 9.

In Table 9, the first column contains the two facial expressions, crying and dazing, of the input sequence. Table 9 shows that 20 of the 435 frames of the dazing expressions are recognized incorrectly. They are classified into the crying class. The frames of crying expressions will be assigned to different incorrect classes. We can observe that the recognition rates of the dazing and crying expressions are 95.40% and 98.15%, respectively.

Table 9. The recognition results of testing sequence 2.

Recognition result Actual facial expressions	crying	dazing	smiling	yawning	sleeping
dazing	20	415			
crying	2876		41	3	10



Fig. 14: Some incorrect examples.

Figure 14 shows some incorrect examples. Figures 14 (a) and (b) show frames of the crying expression classified to the smiling and sleeping classes respectively. Figures 14 (c) shows a frame of the dazing expression classified to the crying class.

## 7 CONCLUSIONS

This paper presents a vision-based infant facial expression recognition system. In this study, a video camera positioned above the infant's crib captures video. The system first detects the infant face by an improve Locus model and extract the feature vectors of infant facial expressions by a PAC method. Finally, the system recognizes the infant facial expression by a probability adaptive model. The experimental results show that the proposed method is robust and efficient.

## ACKNOWLEDGEMENT

The authors would like to thank the National Science Council of the Republic of China, Taiwan for financially supporting this research under Contract No. NSC 99-2221-E-003-019-MY2 and NSC 100-2631-S-003-006-.

## REFERENCES

[1] C. R. De Silvaa, S. Ranganathb, and L. C. De Silvac, "Cloud basis function neural network: Amodified RBF network architecture for holistic facial expression recognition," *Pattern Recognition*, Vol. 41, pp. 1241–1253, 2008.

[2] X. Xie and K. M. Lam, "Facial expression recognition based on shape and texture," *Pattern Recognition*, Vol. 42, pp. 1003–1011, 2009.

[3] W. Gu, C. Xiang, Y. V. Venkatesh, D. Huang, and H. Lin, "Facial expression recognition using radial encoding of local Gabor features and classifier synthesis," *Pattern Recognition*, Vol. 45, pp. 80–91, 2012.

[4] P. Pal, A. N. Iyer, and R. E. Yantorno, "Emotion detection from infant facial expressions and cries," *IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 2, pp. 14–19, 2006.

[5] M. Soriano, B. Martinkauppi, and S. Huovinen, "Skin detection in video under changing illumination conditions," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2002)*, Barcelona, Spain, Vol.1, pp.839-842, 2002.

[6] T. X. Huang, A Smart Digital Surveillance System with Face Tracking and Recognition Capability, Chung Yuan Christian University, Master thesis, 2004.

[7] E. Alpaydin, "Introduction to Machine Learning," *MIT press*, London, 2004.