

AN INFANT FACIAL EXPRESSION RECOGNITION SYSTEM BASED ON MOMENT FEATURE EXTRACTION

C. Y. Fang, H. W. Lin and S. W. Chen

Department of Computer Science and Information Engineering, National Taiwan Normal University, Taipei, Taiwan

Keywords: Facial Expression Recognition, Decision Tree, Moment, Correlation Coefficient.

Abstract: This paper presents a vision-based infant surveillance system utilizing infant facial expression recognition software. In this study, the video camera is set above the crib to capture the infant expression sequences, which are then sent to the surveillance system. The infant face region is segmented based on the skin colour information. Three types of moments, namely Hu, R, and Zernike are then calculated based on the information available from the infant face regions. Since each type of moment in turn contains several different moments, given a single fifteen-frame sequence, the correlation coefficients between two moments of the same type can form the attribute vector of facial expressions. Fifteen infant facial expression classes have been defined in this study. Three decision trees corresponding to each type of moment have been constructed in order to classify these facial expressions. The experimental results show that the proposed method is robust and efficient. The properties of the different types of moments have also been analyzed and discussed.

1 INTRODUCTION

Infants are too weak to protect themselves and lack disposing capacity, and therefore are more likely to sustain unintentional injuries especially when compared to children of other age groups. These incidents are very dangerous and can potentially lead to disabilities and in some cases even death. In Taiwan's Taipei city, the top three causes of infant death are (1) newborns affected by maternal complications during pregnancy, (2) congenital anomalies, and (3) unintentional injuries, which in total account for 83% of all infant mortalities (Doi, 2006). Unintentional injuries are a major cause of infant deaths each year, a majority of which can be easily avoided. Some of the most common causes include dangerous objects surround the infant and unhealthy sleeping environments. Therefore, the promotion of safer homes and better sleeping environments is critical to reducing infant mortality caused by unintentional injuries.

Vision-based surveillance systems, which take advantage of camera technology to improve safety, have been used for infant care (Doi, 2006). The main goal behind the development of vision-based infant care systems is to monitor the infant when they are alone in the crib and to subsequently send warning

messages to the baby-sitters when required, in order to prevent the occurrence of unintentional injuries.

The Department of Health in Taipei city has reported that the two most common causes of unintentional injuries are suffocation and choking (Department of Health, Taipei City Government, 2007). Moreover, in Alaska and the United States, the biggest cause of death among infants due to unintentional injuries is suffocation, which accounts for nearly 65% of all mortalities due to unintentional injuries (The State of Alaska, 2005). The recognition of infant facial expressions such as those when the infant is crying or vomiting may play an important role in the timely detection of infant suffocation. Thus, this paper seeks to address the above problems by presenting a vision-based infant facial expression recognition system for infant safety surveillance.

Many facial expression recognition methods have been proposed recently. However, most of them focus on recognizing facial expressions of adults. Compared to an adult, the exact pose and position of the infant head is difficult to accurately locate or estimate and therefore, very few infant facial expression recognition methods have been proposed to date. Pal *et al.* (Pal, 2006) 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

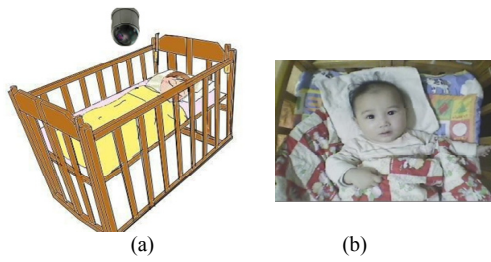


Figure 1: A video camera set above the crib.

facial expressions include anger, pain, sadness, hunger, and fear. The features they used are the local ones. However, we believe that global moments (Zhi, 2008) are more suitable for use in infant facial expression recognition systems.

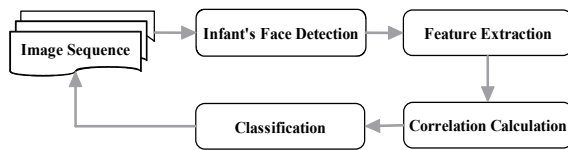


Figure 2: Flowchart of the proposed system.

2 SYSTEM FLOWCHART

The data input to the system consists of video sequences, which have been acquired by a video camera set above the crib as shown in Figure 1(a). An example image taken by the video camera is shown in Figure 1 (b).

Figure 2 shows the flowchart of the infant facial expression recognition system. The system first pre-processes the input image to remove any noise and to reduce the effects of lights and shadows. The infant face region is then segmented based on the skin colour information and then the moment features are extracted from the face region. This study extracts three types of moments as features, including seven Hu moments, ten R moments, and eight Zernike moments.

For each fifteen-frame sequence, the correlation coefficients between two moments (features) of the same type are calculated as the attribute of infant facial expressions. These coefficients aid in the proper classification of the facial expressions. Three decision trees, which correspond to each different type of moment, are used to classify the infant facial expressions.

Five infant facial expressions, including crying, gazing, laughing, yawning and vomiting have been classified in this study. Different positions of the infant head namely front, turn left and turn right have also been considered. Thus, a total of fifteen

classes have been identified.

3 INFANT FACE DETECTION

Three color components from different color models have been used to detect infant skin colour. They are the S component from the HSI model, the Cb component from the $YCrCb$ model and a modified U component from the LUX model. Given a pixel whose colour is represented by (r, g, b) in the RGB color model, its corresponding transfer functions in terms of the above components are:

$$S = 1 - \frac{3}{(r + g + b)} \min(r, g, b) \quad (1)$$

$$Cb = -0.1687r - 0.3313g + 0.5b \quad (2)$$

$$U = \begin{cases} 256 \times \frac{g}{r} & \text{if } \frac{r}{g} < 1.5 \text{ and } r > g > 0, \\ 255 & \text{otherwise.} \end{cases} \quad (3)$$

The ranges of the infant skin colour are defined as $S = [5, 35]$, $Cb = [110, 133]$ and $U = [0, 252]$. These ranges have been obtained from experimental results. Figure 3 (b) shows the skin color detection results of the input image in Figure 3 (a). Figure 3 (c) shows the result after noise reduction and image binarization. Here, a 10×10 median filter has been used to reduce the noise and the largest connected component has been selected as the face region (Figure 3 (d)).

4 FEATURE EXTRACTION

In this section, we will briefly explain the different types of moments. Given an image I , let f represent an image function. For each pair of non-negative integers (p, q) , the digital $(p, q)^{\text{th}}$ moment of I is given by

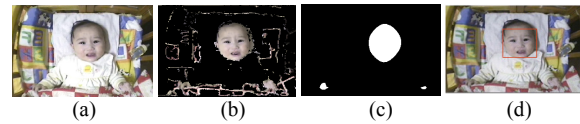


Figure 3: Infant face detection.

$$m_{pq}(I) = \sum_{(x,y) \in I} x^p y^q f(x, y) \quad (4)$$

Let $x_0 = \frac{m_{10}}{m_{00}}$ and $y_0 = \frac{m_{01}}{m_{00}}$. Then the central $(p, q)^{\text{th}}$ moments of I can be defined as

$$\mu_{pq}(I) = \sum_{(x,y) \in I} (x-x_0)^p (y-y_0)^q f(x,y) \quad (5)$$

Hu (Hu, 1962) defined the normalized central moments of I to be

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^\gamma} \quad \text{where } \gamma = \frac{p+q}{2} + 1 \quad (6)$$

From these normalized moments, Hu defined seven moments, which are translation, scale and rotation invariant.

$$\begin{aligned} H_1 &= \eta_{20} + \eta_{02} \\ H_2 &= (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \\ H_3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\ H_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\ H_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\ &\quad + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\ H_6 &= (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\ &\quad + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\ H_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\ &\quad + (3\eta_{12} - \eta_{30})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]. \end{aligned} \quad (7)$$

Liu *et al.* (Liu, 2008) claimed that the Hu moments do not have scale invariability in the discrete case, and therefore proposed ten R moments, which are an improvement over the Hu moment invariants. These R moments can be obtained from the Hu moments as shown below:

$$\begin{aligned} R_1 &= \frac{\sqrt{H_2}}{H_1}, \quad R_2 = \frac{H_1 + \sqrt{H_2}}{H_1 - \sqrt{H_2}}, \quad R_3 = \frac{\sqrt{H_3}}{\sqrt{H_4}}, \\ R_4 &= \frac{\sqrt{H_3}}{\sqrt{|H_5|}}, \quad R_5 = \frac{\sqrt{H_4}}{\sqrt{|H_5|}}, \quad R_6 = \frac{|H_6|}{H_1 \cdot H_3}, \\ R_7 &= \frac{|H_6|}{H_1 \cdot \sqrt{|H_5|}}, \quad R_8 = \frac{|H_6|}{H_3 \cdot \sqrt{H_2}}, \\ R_9 &= \frac{|H_6|}{\sqrt{H_2} \cdot |H_5|}, \quad R_{10} = \frac{|H_5|}{H_3 \cdot H_4} \end{aligned} \quad (8)$$

Zernike moments (Alpaydin, 2004) are defined using polar coordinates and have simple native rotation properties. The kernel of the Zernike moments consists of a set of orthogonal Zernike polynomials defined inside a unit circle. The Zernike moment of order p with repetition q for an image function f is given by

$$Z_{pq} = (C_{pq}^2 + G_{pq}^2)^{1/2} \quad (9)$$

where C_{pq} indicates the real part and G_{pq} indicates the imaginary part and are given by:

$$C_{pq} = \frac{2p+2}{N^2} \sum_{u=1}^{N/2} F_{pq}(2u/N) \sum_{v=1}^{8u} \cos \frac{\pi qv}{4u} f(u,v) \quad (10)$$

$$G_{pq} = \frac{2p+2}{N^2} \sum_{u=1}^{N/2} F_{pq}(2u/N) \sum_{v=1}^{8u} \sin \frac{\pi qv}{4u} f(u,v) \quad (11)$$

where

$$F_{pq}(r) = \sum_{u=0}^{(p-q)/2} (-1)^u \frac{(p-u)!}{u! [p-2u+|q|/2]! [p-2u-|q|/2]!} r^{p-2u}$$

which indicates the radial polynomials and the image size as $N \times N$. For each pixel (x, y) in an image, $u = \max(|x|, |y|)$ and if $u = |x|$, then $v = 2y - \frac{xy}{u}$, otherwise, $v = \frac{2(u-x)y}{|y|} + \frac{xy}{u}$.

As Zernike moments with a larger value of p contain higher frequency information, we select those moments whose value of p is either eight or nine in our experiments. To simplify the index, we use Z_1, Z_2, \dots, Z_{10} to represent $Z_{80}, Z_{82}, \dots, Z_{99}$ respectively.

5 CORRELATION COEFFICIENTS

Given a video sequence $\mathbf{I} = (I_1, I_2, \dots, I_n)$ which describes an infant facial expression, the system can calculate one type of moment for each particular frame. Suppose there are m moments, then the system can obtain m ordered sequences $\mathbf{A}^i = \{A_{i1}, A_{i2}, \dots, A_{in}\}$, $i = 1, 2, \dots, m$, where A_{ik} indicates the i th moment A_i of the frame I_k for $k = 1, 2, \dots, n$. Now the variances of the elements in each sequence \mathbf{A}^i can be calculated by

$$S_{\mathbf{A}^i}^2 = \frac{1}{n-1} \sum_{k=1}^n (A_{ik} - \bar{A}_i)^2 \quad (12)$$

where \bar{A}_i is the mean of the elements in \mathbf{A}^i , and the covariance between \mathbf{A}^i and \mathbf{A}^j is given by

$$S_{\mathbf{A}^i \mathbf{A}^j} = \frac{1}{n-1} \sum_{k=1}^n [(A_{ik} - \bar{A}_i)(A_{jk} - \bar{A}_j)] \quad (13)$$

Therefore, the correlation coefficients between \mathbf{A}^i and \mathbf{A}^j can be defined as


$$r_{\mathbf{A}^i \mathbf{A}^j} = \frac{S_{\mathbf{A}^i \mathbf{A}^j}}{S_{\mathbf{A}^i} S_{\mathbf{A}^j}} \quad (14)$$

Moreover, $r_{\mathbf{A}^i \mathbf{A}^j} = r_{\mathbf{A}^j \mathbf{A}^i}$, $r_{\mathbf{A}^i \mathbf{A}^i} = 1$, for $i, j = 1, 2, \dots, m$. For example, since seven Hu moments have been defined, we can obtain a total of 21 beneficial correlation coefficients. Figure 4 shows a video sequence of an infant crying with fifteen frames. The twenty-one correlation coefficients between the seven ordering sequences are shown in Table 1.



Figure 4: A video sequence of an infant crying.

Table 1: The correlation coefficients between the seven Hu moment sequences.

	H^2	H^3	H^4	H^5	H^6	H^7
H^1	0.1222	0.2588	0.8795	-0.4564	-0.4431	-0.9140
H^2	--	-0.8272	-0.1537	0.6927	-0.1960	0.0573
H^3		--	0.4458	-0.9237	0.2070	-0.3432
H^4			--	-0.6798	-0.2366	-0.9800
H^5				--	-0.1960	0.5663
H^6					--	0.3218

Similarly, we can calculate the correlation coefficients between every two R moments and every two Zernike moments. We believe that the correlation coefficients describe the relationship between these moments, which vary depending on for different facial expressions. Therefore, these coefficients can provide important information, which can be in turn used to classify the different infant facial expressions.

6 CLASSIFICATION TREES

In this study, a decision tree (Alpaydin, 2004), which implements the divide-and-conquer strategy, has been used to classify the infant facial expressions. A decision tree is a hierarchical model used for supervised learning and is composed of various internal decision nodes and terminal leaves. Each decision node implements a split function with discrete outcomes labeling the branches. The advantages of the decision tree are (1) it can perform a quick search of the class of the input features and (2) it can be easily understood and interpreted by mere observation. In this study, we have constructed three binary classification trees corresponding to the three different types of moments.

Suppose K infant facial expressions are to be classified, namely, C_i where $i = 1, \dots, K$. Given a decision node S , let N_S indicate the number of training instances reaching the node S and N_S^i indicate the number of N_S belonging to the class C_i . It is apparent that $\sum_{i=1}^K N_S^i = N_S$. The impurity measure applied in this study is an entropy function

given by

$$E(S) = - \sum_{h=1}^K \frac{N_S^h}{N_S} \log_2 \frac{N_S^h}{N_S} \quad (15)$$

where $0 \log 0 \equiv 0$. The range of this entropy function is $[0, 1]$. If the entropy function is zero, then node S is pure. It means that all the training instances reaching node S belong to the same class. Otherwise, if the entropy is high, it means that the many training instances reaching node S belong to different classes and hence should be split further.

The correlation coefficients $r_{A^i A^j}$ (Eq. (14)) between two attributes A^i and A^j of a training instance can be used to split the training instances. If $r_{A^i A^j} > 0$, then the training instances can be assigned to one branch. Otherwise, the instances can be assigned to a second branch. Let the training instances in S be split into two subsets S_1 and S_2 (where $S_1 \cup S_2 = S$ and $S_1 \cap S_2 = \phi$) by the correlation coefficient $r_{A^i A^j}$. Then the accuracy of the split can be measured by

$$E_{r_{A^i A^j}}(S) = - \sum_{h=1}^K \frac{N_{S_1}^h}{N_{S_1}} \log_2 \frac{N_{S_1}^h}{N_{S_1}} - \sum_{h=1}^K \frac{N_{S_2}^h}{N_{S_2}} \log_2 \frac{N_{S_2}^h}{N_{S_2}}, \quad (16)$$

Finally, the best correlation coefficient selected by the system is

$$r_{A^{i^*} A^{j^*}}(S) = \arg \min_{i,j} E_{r_{A^i A^j}}(S) \quad (17)$$

It is to be noted that once a correlation coefficient has been selected, it cannot be selected again by its descendants.

The algorithm to construct a binary classification tree is shown here:

Algorithm: Decision Tree Construction.

- Step 1: Initially, put all the training instances into root S_R . Regard S_R as an internal decision node and input S_R into a decision node queue.
- Step 2: Select an internal decision node S from the decision node queue. Calculate the entropy of node S using Eq. 15. If the entropy of node S is larger than a threshold T_s , proceed to Step 3, otherwise label node S as a leaf node and proceed to Step 4.
- Step 3: Find the best correlation coefficient $r_{A^{i^*} A^{j^*}}$ to split the training instances in node S using Eqs. 16 and 17. Split the training instances in S into two nodes S_1 and S_2 using the correlation coefficients $r_{A^{i^*} A^{j^*}}$ and then subsequently add S_1 and S_2 into the decision node queue.

Step 4: If the queue is not empty, return to Step 2, otherwise stop the algorithm.

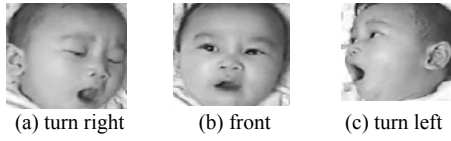


Figure 5: Three head poses of infant yawning.

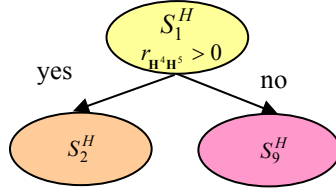


Figure 6: The decision tree of the Hu moments.

7 EXPERIMENTAL RESULTS

The input data for our system was acquired using a SONY TRV-900 video camera mounted above the crib and processed on a PC with an Intel[®]Core™ 21.86GHz CPU. The input video sequences recorded at a rate of 30 frames/second were down-sampled to a rate of six frames/second, which is the processing speed of our current system. In order to increase the processing rate, we further reduced the size (640 x 480 pixels) of each image to 320 x 240 pixels.

Five infant facial expressions, including crying, dazing, laughing, yawning and vomiting have been classified in this study. Three different poses of the infant head, including front, left, and right (an example of an infant yawning as shown from the three positions is shown in Figure 4) have been considered and a total of fifteen classes have been identified.

In the first experiment, the Hu moments and their correlation coefficients were calculated using Eqs. 7 and 14. A corresponding decision tree was constructed using the decision tree construction algorithm. Figure 7 shows the decision tree constructed using the correlation coefficients between the Hu moments as the split function. Node S_1^H is the root, and the split function of S_1^H is $r_{H^4H^5} > 0$. Nodes S_2^H and S_9^H are the left and right branches of S_1^H respectively. The left subtree of the decision tree shown in Figure 6 is illustrated in Figure 7 and the right subtree is depicted in Figure 8.

The split functions of the roots of the left subtree and the right subtree are, $r_{H^3H^5} > 0$ and

$r_{H^6H^7} > 0$ respectively.

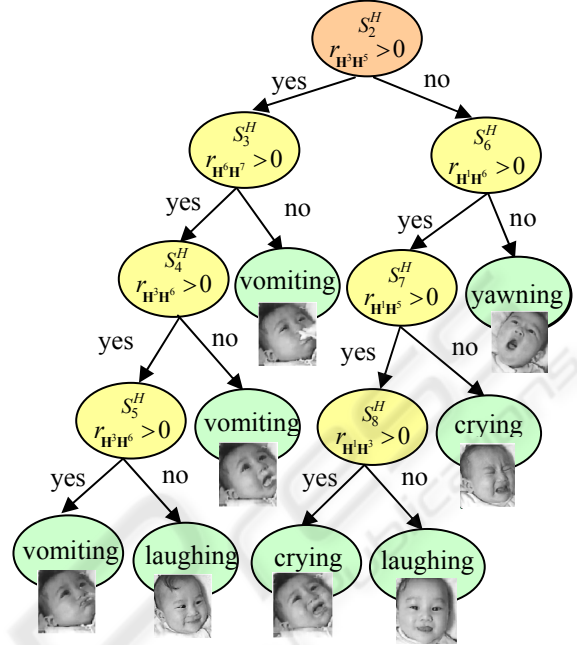


Figure 7: The left subtree of the decision tree depicted in Figure 6.

When Figure 7 and Figure 8 are compared with each other, it can be seen that most of the sequences of the infant head position 'turn right' are classified into the left subtree as shown in Figure 7. Similarly, many sequences of the infant head position 'turn left' are classified into the right subtree as shown in Figure 8.

Similarly, the same fifty-nine fifteen frame sequences were used to train and create the decision trees of the R and Zernike moments. The R moments and their correlation coefficients are calculated using Eqs. 8 and 14. The decision tree created based on the correlation coefficients of the R moments consists of fifteen internal nodes and seventeen leaves with a height of ten. The experimental results are shown in Table 2.

Table 2: The experimental results.

	(1)	(2)	(3)	(4)	(5)
Hu Moments	59	16+17	8	30	90%
R Moments	59	15+17	10	30	80%
Zernike Moments	59	19+20	7	30	87%

- PS. (1) Number of training sequences
 (2) Number of nodes (internal node + leaf)
 (3) Height of the decision tree
 (4) Number of testing sequences
 (5) Classification Rate

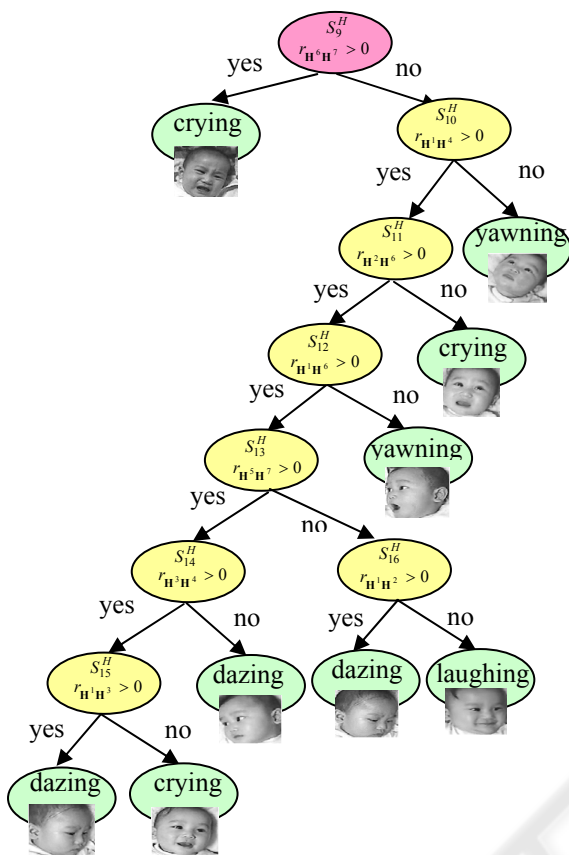


Figure 8: The right subtree of the decision tree shown in Figure 6.

Moreover, the Zernike moments and their correlation coefficients are calculated using Eqs. 9 and 14. The decision tree created based on the correlation coefficients of the Zernike moments includes nineteen internal nodes and twenty leaves, with a height of seven.

Table 2 also shows the classification results of the same thirty testing sequences. We observe that the correlation coefficients of the moments are useful attributes for classifying infant facial expressions. Moreover, the classification tree created from the Hu moments has a smaller height and a fewer number of nodes but a higher classification rate.

8 CONCLUSIONS

This paper presented an infant facial expression recognition technique for a vision-based infant surveillance system. In order to obtain more reliable experimental results, we will be collecting more experimental sequences to construct a more

complete infant facial expression database. Binary classification trees constructed in this study may be less tolerant of. If the correlation coefficients are close to zero, then the noise will greatly affect the results of the classification. The fuzzification of the decision tree may help solve this problem. The infant facial expression recognition system is only one part of the intelligent infant surveillance system. We hope that this recognition system will be embedded into the intelligent infant surveillance system in the near future.

ACKNOWLEDGEMENTS

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 98-2221-E-003-014-MY2 and NSC 98-2631-S-003-002.

REFERENCES

- Doi, M., Inoue, H., Aoki, Y., and Oshiro, O., 2006. Video surveillance system for elderly person living alone by person tracking and fall detection, *IEEE Transactions on Sensors and Micromachines*, Vol. 126, pp.457-463.
- Department of Health, Taipei City Government, 2007. <http://www.health.gov.tw/>.
- The State of Alaska, 2005. Unintentional infant injury in Alaska, *Women's, Children's, and Family Health*, Vol. 1, pp. 1-4, http://www.epi.hss.state.ak.us/mchepi/pubs/facts/fs2005na_v1_18.pdf.
- Pal, P., Iyer, A. N., and Yantorno, R. E., 2006. Emotion detection from infant facial expressions and cries, *IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 2, pp. 14-19.
- Zhi, R., and Ruan, Q., 2008. A comparative study on region-based moments for facial expression recognition, *The Proceedings of Congress on Image and Signal Processing*, Vol. 2, pp. 600-604.
- Hu, M. K., 1962. Visual pattern recognition by moment invariants, *IRE Transactions on Information Theory*, Vol. 8, pp. 179-187.
- Liu, J., Liu, Y., and Yan, C., 2008. Feature extraction technique based on the perceptive invariability, *Proceedings of the Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, Shandong, China, pp. 551-554.
- Alpaydin, E., 2004. *Introduction to Machine Learning*, Chapter 9, MIT Press, Massachusetts, U.S.A.