Face Alive Icons

Xin Li[1], Chieh-Chih Chang[2], Shi-Kuo Chang[1]
[1]University of Pittsburgh, USA,{flying, chang}@cs.pitt.edu
[2]Industrial Technology Research Institute, Taiwan, chieh@itri.org.tw

Abstract

Facial expression is one of the primary communication means of the human. However, realistic facial expression images are not used in popular communication tools on portable devices because of the difficulties in: 1) Acquisition; 2) Transference; 3) Display. In this paper, we propose a system tackling these problems to synthesize facial expression images from photographs for the devices with limited processing power, network bandwidth and display area, which is referred as "LLL" *environment. The facial images are reduced to small-sized* face alive icons *(FAI). Expressions are decomposed into the expression-unrelated* facial features *and the expression-related* expressional features. *The common features are captured and reused across expressions by the discrete model built through the statistical analysis on the training dataset. Semantic synthesis rules are also constructed which reveal the inner relations of expressions. Verified by an experimental prototype system, the approach can produce acceptable facial expressional images utilizing much less computing, network and storage resource than the traditional approaches.* [1]

1 Introduction

Facial expression is one of the primary communication means of the human, and sometimes it is even more expressive than words. Today with the increasing popularity of the advanced communication tools such as emails and short messages, more and more people have been communicating by various means without seeing each other. However, facial expressions are still greatly needed for the most of these conversations.

Observing the popular instant message service (IMS) tools such as the messagers by Yahoo [13] and MSN [7], they use cartoon faces to present the facial expressions such as happiness, sadness, fear, anger, surprise, and so on.

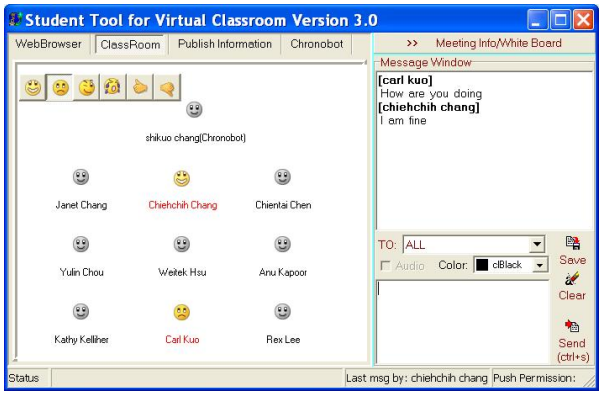

Figure 1. The Student Client of Virtual Classroom System

These cartoon icons become so popular that they are encoded as the combinations of ASCII characters such as ":-)" for happiness and ":-(" for sadness, which are widely used within emails and short messages. The popularity of these cartoon face icons and character combinations has proven the great needs for facial expressions in numerous applications. However, neither of them can provide as natural and vivid expressions as the facial images of the real human.

Our motivation to synthesize realistic facial expressions is originally conceived from the Virtual Classroom (VC) system in *Chronobot* project[1]. VC system simulates a classroom via the Internet, which provides a convenient communication environment for distance learners and teachers like the traditional face-to-face style classroom. As shown in figure 1, user's expressions such as happiness, sadness and so on are represented by the cartoon faces. To display the identifications as well as the expressions, the realistic images are considered to be incorporated. However, for the expression synthesis approaches as far as we surveyed, none of them can fit our needs. In general, the real human's facial images are not used for the expressions in many IMS systems mainly because of the following three reasons:

1. Acquisition: they are hard to acquire. Besides the pri-

[1]Published in the proceeding of the Seventeenth International Conference on Software Engineering and Knowledge Engineering(SEKE'05), Taipei, Taiwan, Jul. 2005.

vacy issue, it costs the users too much effort to take photos for each expression.

2. Transference: they are too big to transfer. The users may use portable devices with the limited processing power and network bandwidth.

3. Display: they are too large to display. There are only small display areas in the portable devices such as cell phones and PDAs.

For the acquisition problem, several approaches are described to synthesize expressions from photographs[12, 11, 5, 8, 10], in which various facial expressions can be generated from one single image. Although these approaches have achieved success in many applications, the proposed algorithms, as far as we studied, are all designed for the high resolution images involving many computing dense operations, which can not be applied directly on the portable devices with limited processing power and network bandwidth.

In this paper, an approach to synthesize facial expression images from photographs is proposed for the portable devices with limited bandwidth, limited processing power and limited display areas, which is called the *"LLL"* environment. The facial images are reduced to the *face alive icons*, which are small size realistic images, usually, 64 pixels by 64 pixels or smaller, so that they are suitable for the most portable devices, and meanwhile still big enough to represent expressions. Based upon the single facial image, the facial features are decomposed into the expression-unrelated features–*facial features*, and the expression-related features–*expressional features*. Through the principal component analysis (PCA) on the training data set, a discrete model for the expressional features can be constructed. The facial alive icons are synthesized by the combination of the standard states of the expression features in the model. The icons can be generated in terminal devices using simple combination rules and operations. The workload of the transference and storage is also reduced. All these make the approach suitable for the portable devices such as cell phones and PDAs.

This paper is organized as follows: the related research is briefly reviewed in Section 2. Comparing with related works, our approach is overviewed in Section 3. Section 4 describes the decomposition rules for expressions. The discrete model of expressional features is explained in Section 5, which is built through the statistical analysis on training data. The synthesis rule and expression encoding are described in Section 6. In Section 7, an experimental system is described, and the result analysis is also presented. The future research is discussed in Section 8.

2 Related Research

A number of approaches have been developed to synthesize facial expressions from photographs [12, 11, 5, 8, 10]. Depending on the basic underlying techniques, generally they can be categorized into two groups: warping-based approaches and morphing-based approaches.

Noh and Neumann[8] have proposed a typical warping-based approach for the expression cloning from one person to another using the vertex motion vectors. The main drawback of this kind of approaches is that it only considers the shape changes, and ignores the other factors such as textures and illuminations which are also important. The morphing-based approaches such as [11, 10] are based on a large collection of the sample expressions. The new expressions can be produced by the morphing between these samples. Although this method could generate the photo-realistic expressions, it does not work for a new person who has no similar samples in the collection.

The work most closely related with ours is a hybrid approach proposed by Wang and Ahuja[12], who decompose face images into three dimensions – the person, expression and feature using High-Order Singular Values Decomposition (HOSVD). Expression synthesis is done by tensor-based transformation in the separate dimensions based upon the training data. It is relatively simple and can produce acceptable result even for new persons. However, as same as all approaches mentioned above, both decomposition and synthesis algorithms require powerful processing capability, and neither are suitable for the portable devices in the *"LLL"* environment.

A compromise solution for the *"LLL"* environment is to perform the complex expression synthesis using high-performance servers, and then distribute the output expression images to low-performance terminal devices. As a result, the terminal devices are only subject to display these images. Although this method can remove the processing workloads of the terminal devices, the heavy burdens for network transference and device storage are still unresolved, which could be a severe problem if a large number of expression images are needed. From this point of the view, our approach is unique because we take expressional features as basic processing units instead of the whole facial images. By combining the different states of the expressional features in the terminal devices, the face alive icons can be generated using much less network and storage resource.

3 Overview of Our Approach

Briefly, our approach contains two distinct steps (Figure 2): the expression decomposition and the icon synthesis. The expression decomposition involves computing

dense operations, so it is performed on high-performance servers. On the other hand, the icon synthesis is a light load process which can be done on the portable devices in the *"LLL"* environment. The connection between the two steps is the user's facial icon profile (FIP), which inclues the facial images and expressional features. The FIP is distributed from the high-performance servers to the portable terminal devices. Since the FIP has much smaller size than a bundle of the expression images, the approach can save a great amount of network bandwidth and device storage resources.

Step	Host	Input	Output
Step1: Expression Decomposition	Servers	Photograph	Facial Icon Profile
Step2: Icon Synthesis	Terminal Devices	Facial Icon Profile	Face Alive Icons

Figure 2. The two steps of our approach

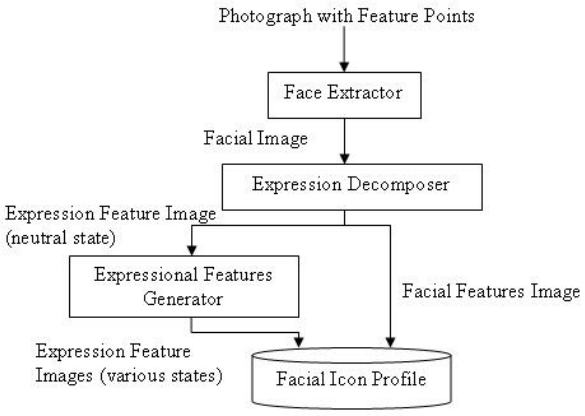

Figure 3. The Process Diagram of the Expression Decomposition

In the step of the expression decomposition,the facial icon profiles are generated from the photograph provided by the users. The process is illustrated in Figure 3. A facial photograph is required to input with feature points such as the eye sockets and lip corners. Using this information, facial images are extracted by the *face extractor* and input to the *expression decomposer*, in which face images are decomposed into the two dimensions: the facial features and expressional features. The facial features are directly copied to the *facial icon profile* because they keep unchanged across expressions. The discrete model of the expressional feature is created in the *expressional features generator* which will be described in following sections. As a result, the facial icon profile contains the facial features as well as the expressional features described by the discrete model. It is compact in size and will be distributed to the terminal devices through network.

The following step is the icon synthesis on the terminal devices. The process is illustrated in Figure 4. The expressions are encoded into uniform formatted codes which defines the states of the expressional features. A synthesis algorithm is proposed using a generation grammar combining the specific states of expressional features to produce the expressions. The operations involved are simply texturing the appropriate expressional features onto the facial features, which is a light load process suitable for portable devices.

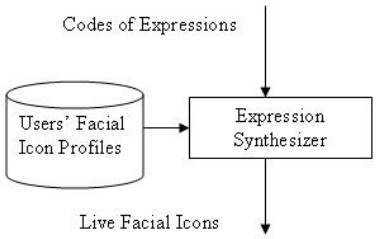

Figure 4. The Process Diagram of the Icon Synthesis

4 Expression Decomposition

Generally, six basic facial expressions are widely recognized by the psychologists, which are happiness, sadness, surprise, anger, disgust and fear[3]. For convenience, we consider the natural expression as the seventh.

To investigate the composition of these typical expressions, a facial image collection – Japanese Female Facial Expressions (JAFFE) database[6] has been used in our approach as the training data set, which contains 210 images of 10 Japanese female subjects. Every image in the database is marked with one major expression according the quantitative evaluations. Figure 5 shows the seven basic expressions of two actresses in the database.

The following two facts are conceived from the observation on different expression images in the JAFFE database:

1. If face can be divided into several independent areas such as eyes area, nose area and so on, some of them are definitely contributing to the expressions much more than the others. For example, in the different expressions, the nose and ears areas usually keep unchanged, while the eyes and mouth area may vary greatly. In the other word, some facial parts are much more expressive than the others.

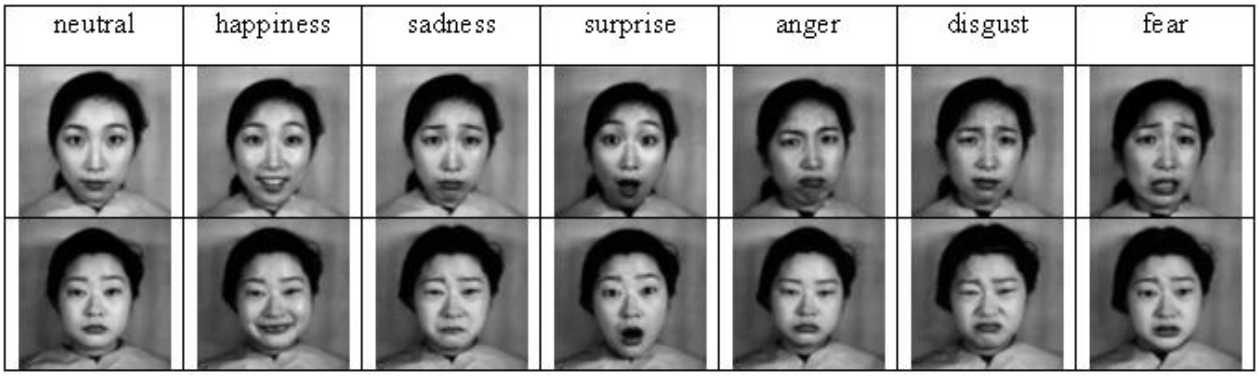

Figure 5. The Seven Basic Expressions in JAFFE database

2. Even for the expressive areas, they may appear similar in different expressions. As shown in Figure 5, the eyes in the expression *sadness* and *disgust* are almost same; the mouths in the expression *angry* and *sadness* are also similar. In the other word, the expressions may share the same appearance of the facial parts; consequently, they could be reused across the expressions.

Based upon the two facts mentioned above, we decompose our face alive icons (FAI) into the following two dimensions:

$$FAI := FF + EF \tag{1}$$

in which:

- FF: Facial Features, are parts of FAI, which are not changing in the different expressions, such as the hair, ears and nose.
- EF: Expressional Features, are parts of FAI, which are changing in the different expressions, such as the eyes and mouth.

It is worth noting that the real human facial image is an extremely complex geometric form, and almost every part of face is active. For example, the human face models used in Pixars Toy Story have several thousands control points each[9]. However, in our approach, we are trying to capture the most expressive parts of face and distinguish them form the relatively inactive parts. In fact, this classification depends on the quality of the expression images required by users. For instance, to produce high-quality photo realistic expressions, we can use as many expressional features as possible. On the contrary, for the facial alive icons in the *"LLL"* environment, the basic requirement is to deliver the expressions as well as the identifications. For this reason, in our approach only the most active facial parts are taken as the expressional features, namely, the mouth and eyes:

$$EF :=< eyes > + < mouth > \tag{2}$$

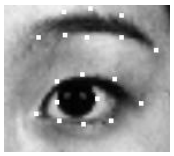

Figure 6. The Eye Area with 18 Landmark Points

Considering some expressions such as winking, two eyes may be in different appearance. So the *eyes* are decomposed into the *left eye* and *right eye*:

$$< eyes >=< left \quad eye > + < right \quad eye > \tag{3}$$

To summarize (1), (2), (3), a grammar for the composition rules of FAI is constructed:

$$FAI := FF + < left \quad eye > + < right \quad eye > + < mouth > \tag{4}$$

5 Discrete Model of the Expressional Features

To eliminate the complexity of the FAI synthesis, we decompose the icons to the expressional features and facial features. However, generally they are still very complex since they could have thousands of subtle appearances for every single person. Here we propose an approach to build a discrete model through Principal Component Analysis (PCA) on the training data, in which *distribution* of the expressional features is investigated, and a reasonable number of *standard states* are defined.

Our discrete model is built similarly with Flexible Model by Lanitis et al. in [4] which has achieved success in many

face modeling applications [2]. Both models use PCA as the main statistical method to capture the main variance of the data; however, besides providing model parameters, the features are "normalized" into the appropriate standard states in the model. What follows is a detail description of the discrete model:

Assuming n sample data items in training set and m variables for landmark points of each item. The i^{th} data item X_i $(i = 1, 2, \cdots, n)$ is:

$$X_i = (x_{i,1}, x_{i,2}, \cdots, x_{i,m}) \tag{5}$$

Where $x_{i,k}$ is the k^{th} variable of the i^{th} data item, and it can be either the coordinates or grayscale of the landmark points.

Based on PCA analysis on the training data, the data item X_i can be assessed as follows:

$$X_i = \overline{X} + P \cdot b \tag{6}$$

In which,

- $\overline{X}$ is the average of the training examples.
- $P = (eig_1, eig_2, \cdots, eig_S), eig_i (i = 1, 2, \ldots, s; s \leq m)$ is the unit eigenvector of the covariance of deviation.
- b is a vector of eigenvector weights referred as Model Parameters.

By modifying b, new instance of model can be generated. Solving b in (6), we get the following equation:

$$b = P^T \cdot (X_i - \overline{X}) \tag{7}$$

Where $P^T \cdot P = 1$.

Usually, the number of eigenvectors needed to describe most of the variability within the training set is much smaller than the original number of variables for landmark points examples e.g. s $\ll$ m. So through this method, features can be described by vector b with much less parameters.

The distance function D is defined on features represented as b_1 and b_2:

$$D(b_1, b_2) = |b_1 - b_2| \tag{8}$$

Assuming there are p expressions $e_1, e_2, \cdots, e_p$. The set of expressions E:

$$E = \{e_1, e_2, \cdots, e_p\}. \tag{9}$$

Categorized by expressions, the vector b of the features for each expression e_i can be represented by the average $\overline{b}_{e_i}$. So we have a basic form of the discrete model S for the feature on the expression set E:

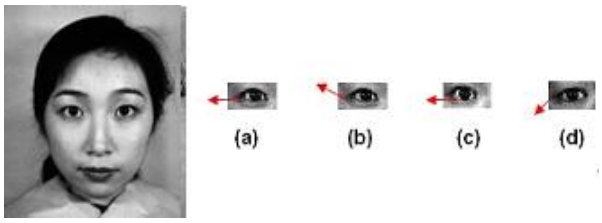

Figure 7. States of the Right Eye: (a) b_1: normal (b) b_2: up (c) b_3: wide-open (d) b_4: down

$$S = \{\overline{b}_{e_1}, \overline{b}_{e_2}, \cdots, \overline{b}_{e_p}\} \tag{10}$$

In which, b_{e_i} is called *standard state* of the feature.

As the fact we discussed in section 4, the features may have similar states in different expressions. An algorithm is proposed to merge similar states in the model S, which is shown in Algorithm 1:

Algorithm 1: Merge items in discrete model S

```
procedure merge(S: Set of b_{e_i})
begin
find b_u, b_v ∈ S which has minimum distance D(b_u, b_v)
    e.g. ∀b_i, b_j ∈ S, D(b_u, b_v) ≤ D(b_i, b_j)
S = S − {b_u, b_v}
b_new = (b_u + b_v)/2
S = S + {b_new}
end;
```

Furthermore, a unique semantic name is also given for each of the standard states according to the appearances. As a result, we have synthesis rules with the semantic meaning followed by the grammar (4), which will be discussed in detail in next section.

For instance, using Algorithm 1, the standard states for the right eye area are merged from seven to four. The images are reconstructed using (6), and the semantic names such as normal, up, wide-open and down are assigned to them according to their appearances, which are shown in Figure 7. Similarly, there are four standard states for the left eye. For the mouth area, there are four standard states which are normal, down-close, up-close and down-open shown in Figure 8:

6 FAI Synthesis and Encoding

Based on the proposed discrete model, the facial icons can be synthesized by the mapping from expressions to the combinations of the standard states of the expressional features. What follows gives a formal description of the synthesis process:

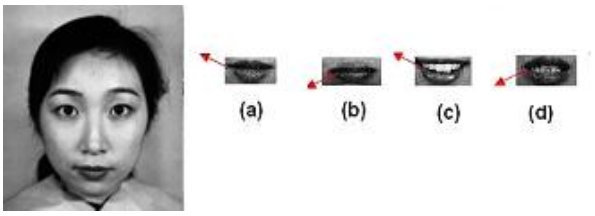

Figure 8. The States of the mouth: (a) b_1: normal (b) b_2: down-close (c) b_3: up-open (d) b_4: down-open

Assuming the expressions are decomposed into N expressional features $f_1, f_2, \cdots, f_N$. For each of facial feature $f_i (i = 1, \ldots, N)$, there are M_i standard states, represented as a set of vectors Ω_i:

$$\Omega_i = \{b_{i,1}, b_{i,2}, \cdots, b_{i,M_i}\} \tag{11}$$

The synthesis rule derived from the grammar (4) can also be described by a function R on the set of the expressions E in (9) to $\Omega_1 \times \Omega_2 \times \cdots \times \Omega_N$, e.g.

$$\forall e \in E, R(e) = (b_{1,k_1}, b_{2,k_2}, \cdots, b_{N,k_N}) \tag{12}$$

in which $b_{i,K_i} \in \Omega_i$.

Given a set of the training images for an expression e, the synthesis rule $R(e)$ can be determined by statistical analysis on the distance functions D in (8) between data item and the standard states. As shown in figure 9, the average distances of the training data with the standard states are listed. The standard states with minimum average distance (numbers underlined) are taken for the synthesis rules followed by the grammar (4) . For example:

$$HAP := FF{+}[Eye : up]{+}[Eye : up]{+}[Mouth : up{-}open] \tag{13}$$

The Rules for seven basic expressions such are listed in Figure 10:

	left eye	right eye	mouth
NEU	normal	normal	normal
HAP	up	up	up-open
SAD	down	down	down-close
SUR	wide-open	wide-open	up-open
ANG	wide-open	wide-open	down-open
DIS	down	down	normal
FEA	wide-open	wide-open	down-close

Figure 10. The synthesis rules

It is worth noting that the synthesis rules can be conceived either from the statistic analysis as we did in Figure 9 or by the intuitive observation on sample expression images. For example, the following non-standard expressions are created by the user's intuition.

$$\begin{aligned} WINKING := \quad & FF + [Eye : down] + [Eye : wideOpen] \\ & + [Mouth : normal] \\ SMILE := \quad & FF + [Eye : normal] + [Eye : normal] \\ & + [Mouth : upOpen] \end{aligned}$$

As discussed in (12), for N expressional features $f_1, f_2, \cdots, f_N$, and $M_i (i = 1, 2, \cdots, N$ standard features for f_i, it needs $\lceil \log_2 M_i \rceil$ bits to encode the feature f_i. Therefore, to encode the expressions with the synthesis rules, totally $\sum_{i=1}^{N} \lceil \log_2 M_i \rceil$ bits are needed.

For example, in our approach, a 2-bits code (e.g. 00, 01, 10, 11) is needed for the states of the each eye and mouth, so totally there are 6 bits for the each expression. Technically, system can support $2^6 = 64$ potential expressions.

7 Experimental System and Result Analysis

An experimental prototype system has been implemented to verify the usability of proposed methodology. Figure 11 shows partial experimental results on two persons. One has same/similar images in the train set, and the other one has not. The method proves to be able to produce acceptable facial icons for the both persons.

Compared with the traditional expression synthesis methodologies [12, 11, 5, 8, 10], in which the facial expressions are considered independent, and each one is stored and transmitted as a image file, our approach reveals the inherit relations among expressions and reuses the expressional features across them. As a result, more expressions can be produced using less storage and network resources. Figure 12 compares the sizes of files needed to transmit and store for displaying $K (K = 7, 15, 30)$ expressions in our approach and the traditional approaches. Considering a communication session involving 20 users, totally around 2.5 MBytes FIPs is needed to be distributed in our approach, which potentially supports up to 64 expressions. For other approaches, however, the total size of image files needed is at lease 19.2 MBytes to support 30 expressions.

8 Discussion and Future Research

In this paper, we propose an approach to synthesize the facial expression images in the *"LLL"* environment. Different from the other methodologies, the facial images are decomposed into expressional features, and the common features are reused across expressions using the discrete model which is built through the statistical analysis on the training data set. The approach has been verified by the experimental prototype system to be able to produce acceptable result utilizing much less network and storage resources.

		NEU	HAP	SAD	SUR	ANG	DIS	FEA
left eye	normal	0.27	1.94	1.65	2.42	3.09	3.32	2.34
	up	2.47	0.45	3.24	2.06	2.21	2.90	2.97
	wideopen	3.00	2.89	2.73	0.75	1.05	2.47	1.30
	down	2.29	3.45	0.96	2.75	4.27	1.23	4.08
right eye	normal	0.33	1.87	1.34	2.36	3.28	3.39	2.35
	up	2.28	0.53	3.57	2.11	2.24	2.99	3.00
	wideopen	2.77	3.08	2.88	0.89	1.11	2.43	1.23
	down	2.06	3.25	0.83	3.07	3.77	1.21	3.92
mouth	normal	0.18	2.35	2.27	3.07	4.01	1.50	2.25
	down-close	2.87	3.45	0.43	2.23	2.25	3.56	1.06
	up-open	3.94	0.94	3.77	0.54	3.17	2.76	3.86
	down-open	3.25	1.95	2.06	2.09	1.23	2.47	3.02

Figure 9. The average distance D distance between training data and the standard states(NEU:neutral; HAP: happiness; SAD:sadness; SUR:surprise; ANG: anger; DIS:digust; FEA:fear).

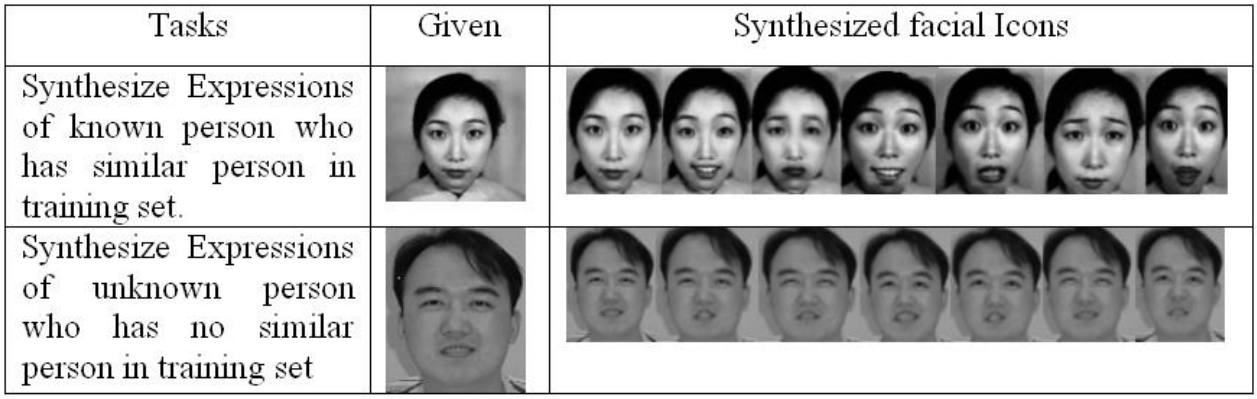

Figure 11. Facial expression synthesis experiments. Facial expression from left to right: neutral, happiness, sadness, surprise, anger, disgust, and fear.

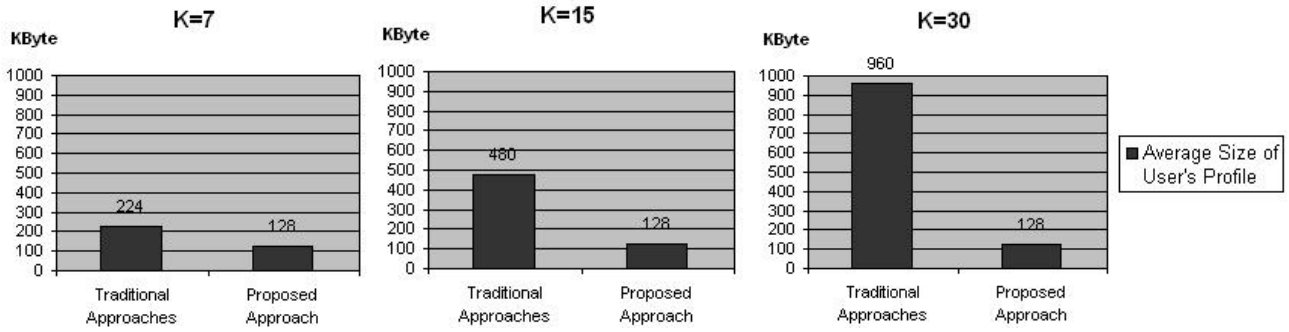

Figure 12. Comparison of the average size of user expression profiles for K expressions per person

It is worth noting that the proposed method is not just for the small-size facial icons. It could be used for the large and high quality expression images: as mentioned in section 4 and 5 , the quality of the output images can be adjusted by the numbers of the expressional features N , the landmark points L and the standard states M. By increasing these

numbers, the quality of the output images can be elevated. However, at the same time the size of facial icon profile (FIP) will also increase. When N, L, M are large enough, the size of FIP will be bigger than the total of separate expression images, and the approach will degenerate to the traditional ones. It is an important and interesting problem to find the optimal trade off between the quality of ouput and the size of FIP. It must be the subject of our future research.

It's also important to point out that the proposed methodology can be extended to new expressions. Recalling that human being can create a new expression by imitating other faces, our system must be feed by the training data to produce new expressions. Based upon the training samples, a new expression can be decomposed using the same way. The discrete model can be updated, and then new synthesis rule can be assessed. As a result, the system could be an open system accepting any new expressions. The further experiment along this line is another subject of our future research.

9 Acknowledgement

This research is a part of *Chronobot* project[1], which is supported in part by the Industry Technology Research Institute(ITRI) and the Institute for Information Industry(III) of Taiwan. We would like to thank Dr. Lyons [6] who kindly provides the JAFFE database, Jui-Hsin Huang for his work on implementation of the experimental prototype system, and Dr. Carl Kuo for the valuable comments.

References

[1] S.K. Chang. A chronobot for time and knowlege exchange. In *Tenth International Conference on Distributed Multimedia Systems*, pages 3–6, San Francisco Bay, CA, Sep. 2004.

[2] Yangzhou Du and Xueyin Lin. Mapping emotional status to facial expressions. In *Proceedings of 16th International Conference on Pattern Recognition*, pages 524–527, Aug. 2002.

[3] P. Ekman. Facial expressions of emotion: an old controversy and new findings. *Philosophical Transactions of the Royal Society*, B335:63–69, 1992.

[4] A. Lantitis, C. J. Taylor, and T. F. Cootes. Automatic interpretation and coding of face images using flexible models. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 19(7):743–756, 1997.

[5] Zicheng Liu, Ying Shan, and Zhengyou Zhang. Expressive expression mapping with ratio images. In *Proceedings of the ACM 28th annual conference on Computer graphics and interactive techniques*, pages 271–276, 2001.

[6] M. J. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings of Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 200–205, 1998.

[7] MSN Messenger. http://messenger.msn.com.

[8] J.Y. Noh and U. Neumann. View morphing. In *SIGGRAPH '01: Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 277–288, 2001.

[9] E. Ostby. Pixar animation studios. *Personal communication*, Jan. 1997.

[10] F. Pighin, J. Hecker, D. Lischinskiy, R. Szeliskiz, and D. H. Salesin. Synthesizing realistic facial expressions from photographs. In *SIGGRAPH '98: Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 75–84, 1998.

[11] S. M. Seize and C. R. Dyer. View morphing. In *SIGGRAPH '96: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 21–30, 1996.

[12] Hongcheng Wang and N. Ahuja. Facial expression decomposition. In *Proceedings of Ninth IEEE International Conference on Computer Vision*, pages 958–965, Oct. 2003.

[13] Yahoo Messenger. http://messenger.yahoo.com/.