

## Short Paper

---

### Automatic Facial Feature Extraction in Model-based Coding\*

MAO-MENG CHUANG, RUEY-FENG CHANG AND YU-LEN HUANG  
*Advanced System Integration Laboratory  
Department of Computer Science and Information Engineering  
National Chung Cheng University  
Chiayi, Taiwan 621, R.O.C.  
E-mail: {rfchang, hyl}@cs.ccu.edu.tw*

Model-based coding is a new image sequence compression technique for very low bit rate coding. Most researchers have paid more attention to facial image analysis and synthesis because facial images are very important in multimedia communication applications, such as video phone, video conferencing, remote learning etc. This coding method represents the image content in a structural way. This is the reason why model-based coding can get a higher compression ratio (1-10 kb/s) than can the conventional coding methods. In order to encode image signals efficiently, it is necessary to create a suitable generic model and adapt it to the actual object accurately. In this paper, we propose a scheme, based on the integral-projection algorithm, which adapts a face model to an actual face automatically. First, the image is preprocessed by means of edge detection. From the edge mapped image, we use the feature of the increasing sudden edge density to indicate the rough vertical positions of eyes and mouth. According to the local threshold value in each feature area, all of the control points about the eyes and mouth are found. Finally, we use these control points to adapt the model to an actual face.

**Keywords:** facial model, facial features, integral-projection algorithm, model-based coding, very low bit rate coding

#### 1. INTRODUCTION

Recently, model-based [1-8] coding has become a popular research topic in image communication. This method encodes an image at a very low bit rate while retaining a certain level of quality. Generally, the model-based coding scheme has two parts, image analysis and image synthesis [3, 6, 7], that are based on a 3-D model and knowledge and experience of the object in the image. Since facial images are very important for many applications, including video phone and video conferencing, the model-based coding method has concentrated on facial image analysis and synthesis [1, 2, 4, 8-10]. For CIF or QCIF color head-and-shoulder sequences, reproductions can be obtained at very low bit rates of 16-64 kb/s.

---

Received November 8, 1997; revised May 2 & July 10, 1998; accepted October 8, 1998.  
Communicated by Jhing-Fa Wang.

\*This work was supported by the National Science Council, Taiwan, R.O.C., under Grant NSC85-2213-E-194-016.

The key tasks points of model-based coding are creating a facial model and adapting it to an actual face. Thus, the analysis and synthesis methods are based on a good facial model. In most of the model-based analysis and synthesis image coding (MBASIC) systems, the facial feature points are extracted manually. In order to make model-based coding practical, we propose in this paper a scheme to extract these feature points and to adapt the model automatically. In our scheme, we first use histogram equalization, Gaussian, and Sobel operators to preprocess images and to find all the edge points. The face area is found, and only the edge points in the face area are retained. The integral-projection algorithm [11] is used for the facial edge points to find the facial feature positions. Then, using the thresholding operation with a different local threshold value in each facial feature area, every control point about the facial features, such as the eyes and mouth, are found. Finally, we use these control points to adapt the model to an actual face automatically.

## 2. AUTOMATIC FACIAL FEATURE EXTRACTION AND MODEL ADAPTATION SCHEME

In the analysis process in model-based coding, automatic facial model adaptation is the key procedure. Many methods for model adaptation have been proposed [8, 12-18]. However, there are some constraints on finding and tracking facial features. Hence, a new scheme based on edge detection and the edge pixel number is proposed in this paper. It is called automatic facial feature extraction and the model adaptation (AFFEMA) scheme. The proposed scheme contains four parts: facial area extraction in an entire image, estimation of the vertical positions of the facial features, such as the eyes and mouth, extraction of control points of facial features, and facial model adaptation. Automatic facial model adaptation is shown in Fig. 1.

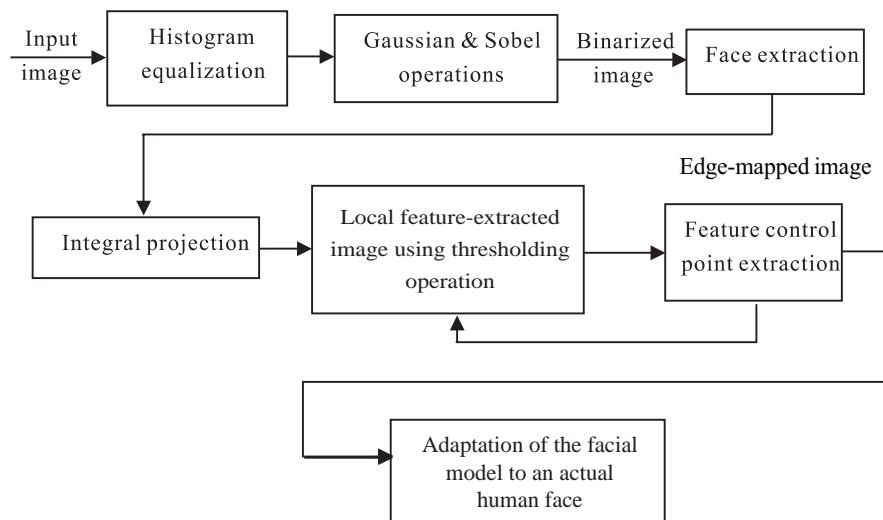


Fig. 1. The AFFEMA scheme.

## 2.1 Facial Area Extraction

In the AFFEMA scheme, the first step is to find the face location and separate it from the background. Before locating the face area, a preprocessing procedure is necessary to alleviate noise and get a good result from edge detection. In our scheme, histogram equalization and the gaussian operator are used for this task. Because histogram equalization can produce an image with higher contrast, it can enhance the edge continuity of an image. The Gaussian operator will blur the image in order to reduce noise. Because edge distribution is important for the next task of facial area searching, and noise always creates some false edges, the Gaussian operator is used to reduce the noise effect. Then, we convolve the blurred image using vertical and horizontal Sobel edge operators to get a binarized edge image.

From the binarized edge image, we can separate the head from the background using the following search scheme. First, for each horizontal line, we find the two edge points which are close to both borders of the image. By using the location information of these edge pixels and finding the brighter pixels, the area which includes the head and shoulder can be generated. Because our feature searching scheme focuses on the facial part, we must remove the hair. In general, the gray level of hair is lower than that of a face, and the height of the head area is about two-thirds the height of the total area, which includes the head and shoulder. In addition lighting is an important factor in extraction of the face because it may destroy the integrity of the face. In order to conquer this lighting problem, the threshold values, which include luminance and chrominance, are estimated from the area of the head and are used to separate the hair. That is, if a pixel have either a larger luminance value or a larger chrominance value, then it is considered to be a pixel in the head area. Then, we use the binarized image from which the entire hair area has been removed to locate the facial part. We find the two lowermost side pixels whose gray levels are larger than the threshold values because the the face is assumed to be brighter. According to the positions of these two pixels, a bottom-up searching process that finds the side pixels of the facial part is performed, and a new area that includes only the face and neck (or clothes) is generated (See Fig. 2). Finally, only the edge pixels in this new area are retained, and others are discarded. We can obtain an edge map image  $I_{edge}(x, y)$  that can be used to estimate the face feature locations.

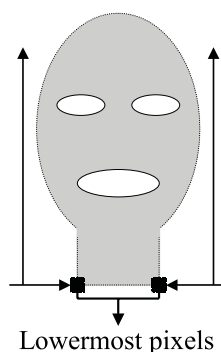


Fig. 2. Illustration of the searching method for the facial area.

## 2.2 Estimation of the Vertical Positions of Facial Features

A mathematical statistical method is used to estimate the approximate vertical positions of the facial features, such as the eyes and mouth, in our AFFEMA scheme. The integral projection, a technique that was originally proposed by Kanade and was modified by De Silva *et al.* [11], is applied to our problem of facial feature position estimation. In [11], the integral projection was used to find the eye plane. However, in our method, the integral projection is used to first find the nose position and then the eye and mouth positions. An edge map image  $I_{edge}(x, y)$  is divided into some areas of equal size in the vertical direction. Then, the vertical integral projection of the edge map image in a rectangular area of  $1 \leq x \leq IW$  and  $i * step + 1 \leq y \leq (i + 1) * step$  is estimated by

$$v(i, x) = \sum_{y=i*step+1}^{(i+1)*step} I_{edge}(x, y),$$

where  $v(i, x)$  indicates that the number of the edge pixels which have the same  $x$  coordinate in the  $i$ th rectangle,  $IW$  is the width of the input image, and  $step$  is the rectangle height, where the image height is divided by  $N$ .  $N$  is the number of equal size rectangles. In our experiment,  $N$  was 36 and 72 for the image sequences Miss America and Claire, respectively. Then all  $v(i, x)$ , for  $1 \leq x \leq IW$  are summed up, and this summation indicates the number of edge pixels in the rectangular area  $i$ . When searching from the top of the head, we can observe a sudden increase in the number of edge pixels in the eye-plane. In this paper, we call this sudden increase the peak. In order to keep a forehead covered by hair from influencing the peak decision, we first find the peak at the nose. According to the edge pixel number, the vertical positions of the eyes and mouth are estimated by means of upward and downward searching, respectively.

After the vertical positions of the face features have been found, we calculate two distances to locate these feature areas roughly. These distances are  $d_{eye}$  and  $d_{mouth}$ .  $d_{eye}$  is the distance from the middle of eyes to the nose.  $d_{mouth}$  is the distance from the mouth to the midpoint between the nose and mouth. This is shown in Fig. 3.

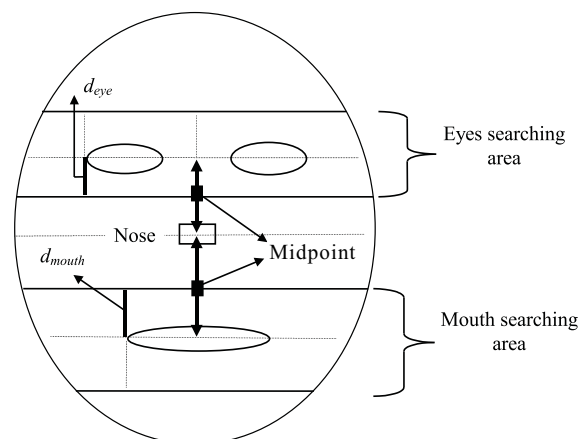


Fig. 3. Illustration of  $d_{eye}$ ,  $d_{mouth}$ , and the location area about each feature.

### 2.3 Extraction of Feature Control Points

Based on the two feature areas (eyes and mouth), we can find the control points which reflect the variation of the feature shapes. The thresholding operation is used to extract the control points. A mean value for each eye area is calculated and taken as a threshold value. Four control points (E1-E4), as shown in Fig. 4, are found for the left and the right eyes. The lowermost point (E1) and the point near the ear (E2) are first detected by means of bottom-up searching. The one near the nose (E3) is obtained along the line constructed by the two control points E2 which have been found. The search direction for E3 is from the center of eyes to the side of the face (see Fig. 5).

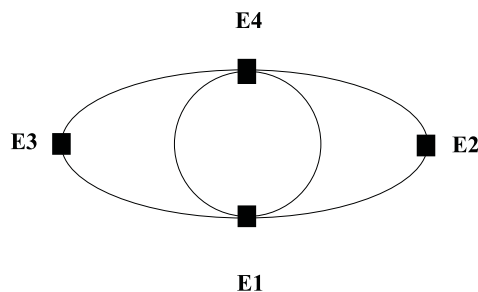


Fig. 4. Four control points of the left eye.

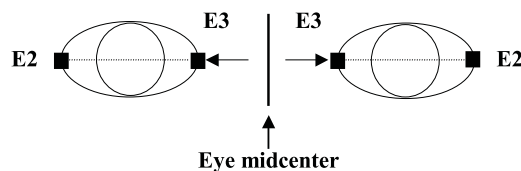


Fig. 5. The search scheme for control point E3.

The last one is the uppermost point of an eye. The searching process is more complicated than that for the others because of the eyebrows. In order to reduce the possibility of mistaking the upper side of the eyelid for the upper edge of the eyes, we adopt the following procedure. We first consider a line upward extending from the center of the eyes orthogonal with the line connecting the left and right control points. A bright pixel is searched for on the vertical line from bottom to top in the upper area of the eye. We then take the position of the bright pixel as the uppermost position of the eye.

In the detection process, lighting is a key factor affecting our result because it may generate some unexpected shadows near facial features. In order to solve this problem, we refer to not only the luminance information  $Y$ , but also the chrominance information  $V$  in the thresholding operation. That is, if the luminance value of a pixel is larger than a luminance threshold, or if its chrominance value is larger than a chrominance threshold, then it is assigned as a bright pixel. The improvement obtained using the chrominance information is shown in Fig. 6.



Fig. 6. The result showing improvement for the left eye in frame 1 of the sequence Miss America. (a) Using only the luminance information. (b) Using both of the luminance and chrominance information.

Like the eyes, the basic thresholding operation with a predetermined threshold value is used in the detection of control points about the mouth, and these four points (M1-M4) shown in Fig. 7 are found in order to represent the change in shape. First, we search for the extreme positions of dark pixels as the mouth control points M1 and M2 from both sides to the center of the face. Based on these two points, we can find the lowermost mouth control point M3. The searching direction is along a line  $L$  from the bottom of the mouth feature area to the top. This line is located at the midpoint between the two control points M1 and M2 and is orthogonal with the line connecting these two control points M1 and M2. The uppermost point is detected using the bottom-up searching process from both sides to the center of the face. When the uppermost point is found, we use the line  $L$  again to do top-down searching from the vertical position of the uppermost point, and then the control point M4 is located. The search method used to find the control points M3 and M4 is shown Fig. 8.

Moreover, we propose an adaptive gray-level thresholding method to improve the stability of our searching scheme. Several constraints are used to decide whether the threshold should be adjusted. Note that we only adjust the luminance threshold because the luminance information is more important than the chrominance information. If the following situations occur, the luminance threshold will be changed. (1) One of these feature control points can not be found. (2) The vertical distance between the upper and lower control points is larger than the horizontal distance between the left and right control points in the eyes. (3) The horizontal distance between the left and right control points is more than two times the distance between the upper and lower control points in the mouth.

## 2.4 Facial Model Adaptation

In our AFFEMA scheme, we use the wire frame model CANDIDE [7] as our object model. It is composed of 64 points and 96 triangles of different sizes. Our goal is to adapt the model to an actual face using the control points we have estimated previously. Because we only get 2-D information for an input image, the adaptation process is also based on the 2-D plane.

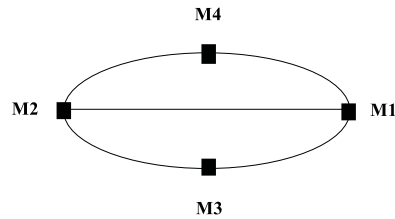


Fig. 7. Four control points of the mouth.

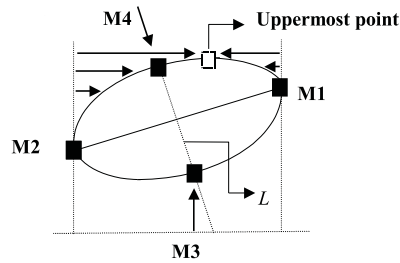


Fig. 8. Illustration of the search method for the control points M3 and M4.

The adaptation of a facial model proceeds in two successive steps: (1) global adaptation and (2) local adaptation. Global adaptation accounts for changing the size of the wireframe model and its position to fit the facial contour of the input image. Local adaptation deals with changing the shapes of facial features and applying the control points which were obtained previously to reflect the shape variation of the eyes and mouth in the model.

Global adaptation of the facial model consists of scale, translation, and rotation of the face. The details of these operations are as follows. (1) Scale: We first estimate the distance between the control points E2 of the left and right eyes in the image. A scale factor of the  $x$  coordinate is the ratio of the distance between the two control points E2 in the actual face to the corresponding distance in the facial model shown in Fig. 9. Like the above step, we estimate the distance between the midpoint between the of two eyes and the center of the

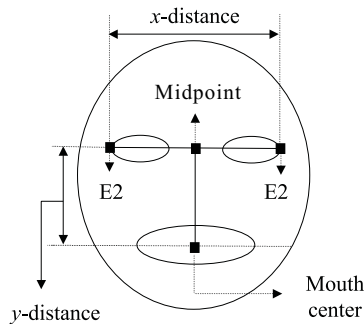


Fig. 9. Illustration of the referenced distances for the scale factors in the  $x, y$  directions.

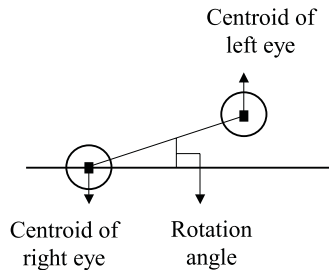


Fig. 10. Illustration of head inclination.

mouth. The scale factor of the  $y$  coordinate is the ratio of this distance to the related distance in the facial model. Using these two factors, the facial model is scaled up or down to fit the actual face. (2) Translation: The translation operation is performed using the midpoint between the eyes as a reference coordinate. From the reference coordinate, an offset between the model and the actual face can be estimated, and the position of the model can be shifted to the actual position of the face in the image. (3) Rotation: The inclination of the head is estimated as the angle between the two eyes and a horizontal line as shown in Fig. 10. We find the centroids of the left and right eyes, respectively, and a vector then is formed by these two centroids. Using the vector, we estimate the angle  $\theta$  by means of the mathematical formula arc sine.

The goal of local adaptation is for the adaptation to express the motion of the facial features and to provide a more refined fit in the actual face. The implementation of this adaptation is based on the control points which have found in previous steps. Using the control point information, the local deformation of this facial model is obtained by adjusting the positions of the control points in the model to match the coordinates of the corresponding control points in the actual face. That is, the actual feature control points will directly replace the corresponding points in the facial model.

### 3. SIMULATION RESULTS

An excellent feature extraction method must satisfy three requirements: it must perform automatic extraction, it must be user independent, and it must be environment independent [8]. But in most cases, because the chrominance and luminance of the background around the user's head are close to those of the human face, separation of the face from the background becomes more difficult. On the other hand, the pose of the user and objects other than organs, such as glasses in the face, also influence the stability of feature extraction. To reduce the complexity of the problem, we make some reasonable assumptions for the input image. (1) The person faces the camera, and the inclination of the head is less than 20-30 degrees. (2) The background is smooth, and its luminance is different from that of the head and clothes of the person. (3) The gray level of the person's hair is lower than that of the person's face. (4) The person does not wear glasses and has no beard or mustache.

In the computer simulation, we used two test image sequences, Miss America and Claire, which satisfy the assumptions listed above with a size of 352x288 and 256 gray levels. The number of test frames was 90. In this work, human observation was used to determine the performance of the simulation results. Our criteria were as follows:

1. An eye or mouth was considered said to be located correctly if the vertical position that we estimated matched the vertical position of this feature.
2. The position of each control point was near the contour of the corresponding feature.

Based on these criteria, the simulation performance of feature location and control point searching are summarized in Table 1. From our experimental results and the assumptions, we find that the positions of the face features were usually found correctly in the input images. And for the sequence Miss America, we also conquer that the left eye was affected by shadow. The luminance and chrominance information was used to remove the shadow effect and get a good result. For these two test sequences, the overall percentage of correctness in finding the control point positions was 94.5%. We also show different results for various eye (see Fig. 11) and mouths (see Fig. 12) samples for these two test image sequences used in this simulation. Finally, results of the adaptation of the facial models to these samples are shown in Fig. 13.

**Table 1. The result for processed frames in terms of feature locations and control point positions.**

Sequence	Number of test frames	Number of error frames for feature locations	Number of error frames for control point positions	Number of correct frames
Miss America	90	0(0%)	3(3%)	87(97%)
Claire	90	3(3%)	4(4%)	83(92%)



Fig. 11. Various examples of processed results for the control points of the eyes in the sequence Miss America. (a) 7th frame. (b) 9th frame. (c) 39th frame.



Fig. 12. Various examples of processed results for the control points of the mouth in the sequences Miss America. (a) 19th frame. (b) 35th frame. (c) 43th frame.

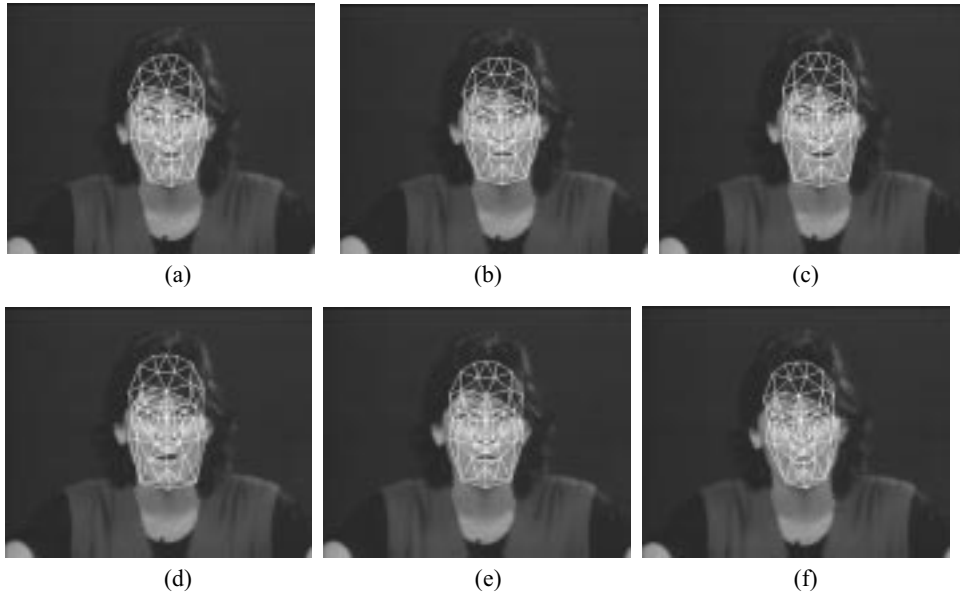


Fig. 13. The result of adaptation of the facial models to the sequence Miss America. (a) 7th frame. (b) 9th frame. (c) 39th frame. (d) 19th frame. (e) 35th frame. (f) 43th frame.

#### 4. CONCLUSIONS

In this paper, we have proposed a feature extraction and model adaptation method that is called AFFEMA. The scheme is based on a statistical method-integral projection and edge detection used to locate each face feature position. Although our scheme has with several limitations, such as background color and facial objects, some specific contributions to feature extraction are provided by this scheme. (1) The user's face does not need to appear at a fixed position in the input image. (2) The area of the person's face does not need to be fixed. In other words, the distance from the user to the camera can be varied. (3) No

parameter need be predetermined except for the *step* parameter. Any threshold or parameter, such as the threshold value of luminance or chrominance, and the size of the area for each feature are adaptive. (4) The geometry of the eye pair region, and that of the mouth region is not limited. Besides the fact that it can be combined with the model-based coding system for practical use, our AFFEMA scheme can also be extended in other ways, such as to recognition and identification of human faces. It provides a simpler and more reliable solution to the feature point extraction problem.

## REFERENCES

1. M. Kaneko, A. Koike, and Y. Hatori, "Coding of facial image sequence based on a 3-D model of the head and motion detection," *Journal of Visual Communication and Image Representation*, Vol. 2, No. 1, 1991, pp. 39-54.
2. T. Fukuhara and T. Murakami, "3-D motion estimation of human head for model-based image coding," *IEE Proceedings I- Communications, Speech, and Vision*, Vol. 140, 1993, pp. 26-35.
3. K. Aizawa and T. S. Huang, "Model-based image coding: Advanced video coding techniques for very low bit-rate applications," *Proceedings of the IEEE*, Vol. 83, No. 2, 1995, pp. 259-271.
4. H. Li, P. Roivainen, and R. Forchheimer, "3-D motion estimation in model-based facial image coding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 15, No. 6, 1993, pp. 545-555.
5. G. Bozdag, A. M. Tekalp, and L. Onural, "3-D motion estimation and wireframe adaptation including photometric effects for model-based coding of facial image sequences," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 4, No. 3, 1994, pp. 246-256.
6. C. S. Choi, K. Aizawa, H. Harashima, and T. Takebe, "Analysis and synthesis of facial image sequences in model-based image coding," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 4, No. 3, 1994, pp. 257-275.
7. M. J. T. Reinders, P. J. L. van Beek, B. Sankur, and J. C. A. van der Lubbe, "Facial feature localization and adaptation of a generic face model-based coding," *Signal Processing: Image Communication*, Vol. 7, No. 1, 1995, pp. 57-74.
8. H. C. Huang, M. Ouhyoung, and J. L. Wu, "Automatic feature point extraction on a human face in model-based image coding," *Optical Engineering*, Vol. 32, No. 7, 1993, pp. 1571-1580.
9. S. C. Pei and M. S. Su, "An interactive tool for synthesis of facial expression based on three-dimension model," in *8th Chinese Image Processing and Pattern Recognition Society Conference on Computer Vision, Graphics and Image Processing*, 1995, pp. 388-394.
10. C. L. Huang and C. W. Chen, "Human facial feature extraction for face interpretation and recognition," in *11th International Association for Pattern Recognition Conference on Pattern Recognition, Conference B: Pattern Recognition Methodology and Systems*, Vol. II, 1992, pp. 204-207.
11. C. de Silva, K. Aizawa, and M. Hatori, "Detection and tracking of facial features," in *SPIE Visual Communications and Image Processing*, Vol. 2501, 1995, pp. 1161-1172.

12. R. Brunelli and T. Poggio, "Face recognition: Features versus templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 15, No. 10, 1993, pp. 1042-1052.
13. A. L. Yuille, P. W. Hallinan, and D. S. Cohen, "Feature extraction from faces using deformable templates," *International Journal of Computer Vision*, Vol. 8, No. 2, 1992, pp. 99-111.
14. P. J. L. van Beek, B. Sankur, and J. C. A. van der Lubbe, "Contour extraction and matching for image sequence coding," in *IEEE International of Conference on Image Processing and its Applications*, 1992, pp. 109-114.
15. R. Brunelli and T. Poggio, "Face recognition through geometrical features," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1992, pp. 792-800.
16. I. Craw, D. Tock, and A. Bennett, "Finding face features," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1992, pp. 92-96.
17. O. Nakamura, S. Mathur, and T. Minami, "Identification of human faces based on isodensity maps," *Pattern Recognition*, Vol. 24, No. 3, 1991, pp. 263-272.
18. M. J. T. Reinders, B. Sankur, and J. C. A. van der Lubbe, "Transformation of a general 3D facial model to an actual scene face," in *11th International Association for Pattern Recognition Conference on Pattern Recognition, Conference C: Image, Speech and Signal Analysis*, Vol. III, 1992, pp. 75-78.

**Mao-Meng Chuang** (莊茂盟) was born in Taipei, Taiwan, on September 5, 1968. He received the B.S. degree in chemical engineering from Tamkang University, Taiwan, Republic of China, in 1991, and the M.S. degree in computer science and information engineering from National Chung Cheng University, Taiwan, Republic of China, in 1996. He is currently a high-level office worker in the Department of Information, Union Bank, Taiwan, Republic of China. His research interests include still image coding, video sequence coding, and computer networking.

**Ruey-Feng Chang** (張瑞峰) was born in Taichung, Taiwan, on August 25, 1962. He received the B.S. degree in electrical engineering from National Cheng Kung University, Taiwan, Republic of China, in 1984, the M.S. degree in computer and decision sciences and the Ph.D. degree in computer science from National Tsing Hua University, Taiwan, Republic of China, in 1988 and 1992, respectively. He is currently an Associate Professor in the Department of Computer Science and Information Engineering, National Chung Cheng University, Taiwan, Republic of China. His research interests include still image coding, video sequence coding, and packet video. Dr. Chang is a member of IEEE, ACM, SPIE, and Phi Tau Phi.

**Yu-Len Huang** (黃育仁) was born in Chiayi, Taiwan, on May 22, 1970. He received the B.S. degree in computer science from Tung Hai University, Taiwan, Republic of China, in 1990, and the M.S. and Ph.D. degrees in computer science and information engineering from National Chung Cheng University, Taiwan, Republic of China, in 1994 and 1999. His research interests include still image coding, video sequence coding, neural networking, computer networking, fuzzy system, and packet video.