# 다양한 형식의 얼굴정보와 준원근 카메라 모델해석을 이용한 얼굴 특징점 및 움직임 복원

김 상 훈†


## 요 약

본 논문은 MPEG4 SNHC의 얼굴 모델 인코딩을 구현하기 위하여 연속된 2차원 영상으로부터 얼굴영역을 검출하고, 얼굴의 특징데이터들을 추출한 후, 얼굴의 3차원 모양 및 움직임 정보를 복원하는 알고리즘과 결과를 제시한다. 얼굴 영역 검출을 위해서 영상의 거리, 피부색상, 움직임 색상정보등을 융합시킨 멀티모달합성의 방법이 사용되었다. 결정된 얼굴영역에서는 MPEG4의 FDP(Face Definition Parameter)에서 제시된 특징점 위치중 23개의 주요 얼굴 특징점을 추출하며 추출성능을 향상시키기 위하여 GSCD(Generalized Skin Color Distribution), BWCD(Black and White Color Distribution)등의 움직임색상 변환기법과 형태연산 방법이 제시되었다. 추출된 2차원 얼굴 특징점들로부터 얼굴의 3차원 모양, 움직임 정보를 복원하기 위하여 준원근 카메라 모델을 적용하여 SVD(Singular Value Decomposition)에 의한 인수분해연산을 수행하였다. 본 논문에서 제시된 방법들의 성능을 객관적으로 평가하기 위하여 크기와 위치가 알려진 3차원 물체에 대해 실험을 행하였으며, 복원된 얼굴의 움직임 정보는 MPEG4 FAP(Face Animation Parameter)로 변환된 후, 인터넷상에서 확인이 가능한 가상얼굴모델에 인코딩되어 실제 얼굴과 일치하는 모습을 확인하였다.


# Facial Features and Motion Recovery using multi-modal information and Paraperspective Camera Model

Sang-Hoon Kim†


## ABSTRACT

Robust extraction of 3D facial features and global motion information from 2D image sequence for the MPEG-4 SNHC face model encoding is described. The facial regions are detected from image sequence using multi-modal fusion technique that combines range, color and motion information. 23 facial features among the MPEG-4 FDP (Face Definition Parameters) are extracted automatically inside the facial region using color transform (GSCD, BWCD) and morphological processing. The extracted facial features are used to recover the 3D shape and global motion of the object using paraperspective camera model and SVD (Singular Value Decomposition) factorization method. A 3D synthetic object is designed and tested to show the performance of proposed algorithm. The recovered 3D motion information is transformed into global motion parameters of FAP (Face Animation Parameters) of the MPEG-4 to synchronize a generic face model with a real face.


**키워드 : 얼굴검출(face detection), 피부 고유색 변화(skin-color transform), 거리정보분할(range segmentation), FAP, 준원근 카메라 모델(paraperspective camera model)**

## 1. Introduction

Recently, the focus on video coding technology has shifted to real-time object-based coding at rates of 8kb/s or lower. To meet such specification, MPEG-4 standardizes the coding of 2D/3D Audio/Visual hybrid data from natural and synthetic sources [1]. Specifically, to provide synthetic image capabilities, MPEG-4 SNHC AHG[13] has focused on the interactive/synthetic model hybrid encoding in the virtual space using real facial object perception technique. Although face detection is a prerequisite for the facial feature extraction, still too many assumptions are required. Generally, informations of range, color and motion of human object have been used independently to detect faces based on the human perception technique. The range information alone is not enough to separate a general facial object from background due to its frequent motion.

Also the motion information is not sufficient to distinguish

※ 본 연구는 한경대학교 2002년도 학술연구조성비의 지원에 의한 것임.
† 정 회 원 : 국립한경대학교 제어계측공학과 교수
논문접수 : 2002년 7월 27일, 심사완료 : 2002년 10월 26일

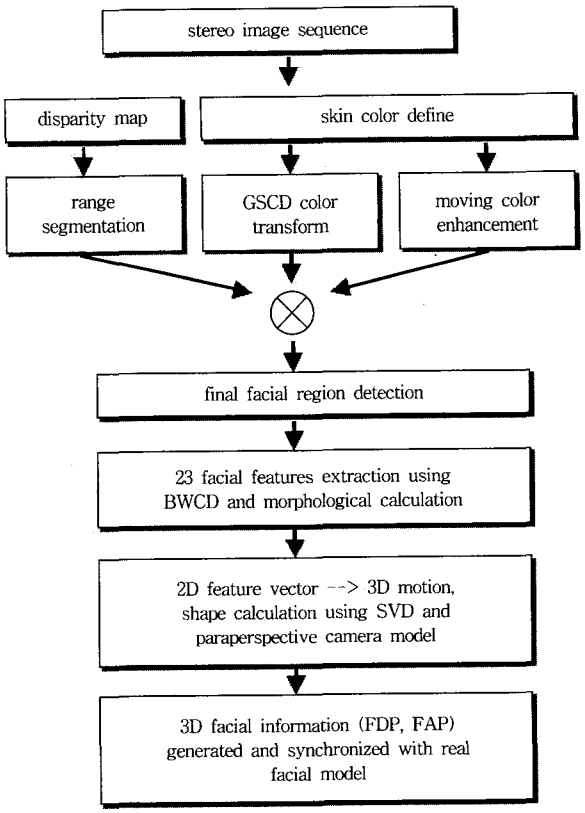
a facial object from others. Recently, facial colors have been regarded as a remarkable clue to distinguish human objects from other regions [9]. But facial color has still many problems like similar noises in the complex background and color variation under the lighting conditions. Recent papers have suggested algorithms that combine range and color information [11], and that use multi-modal fusion range, color and motion information [15] to detect faces exactly. Two approaches are in use generally to recover object's motion and shape information, such as information flow-based method [2, 3] and feature-based method [4-6]. The information flow-based method uses plane model [2] or oval face model [3], which accumulates errors as the number of image frame is increasing. The feature based method has a problem of diverging to an uncertain range according to the initial values [4, 5], or works only under the assumption that the 3D structure of objects is already known [6]. The factorization-based method using paraperspective camera model shows the robustness of recovering 3D information. The batch-type SVD computation method [7] is enhanced further to operate sequentially for the real-time execution [8]. However, it still has a problem in extracting features automatically and has a large error in calculating a depth information.

The goal of this paper is

1) robust extraction of the 3D facial features from image sequence with complex background, and
2) recovery of the global motion information from the extracted facial features for the MPEG-4 SNHC face model encoding.

Facial regions are detected from a image sequence using a multi-modal fusion technique that combines range, color and motion information. The 23 facial features suggested by the MPEG-4 FDP (Face Definition Parameters) are extracted automatically.

The extracted facial features are used to recover the object's 3D shape and global motion sequentially based on the paraperspective factorization method. The proposed algorithm recover depth estimation error of the single paraperspective camera model. Finally, the recovered facial 3D motion and shape information is transformed into the global FAP motion parameters of the MPEG-4 to handle the synthetic face model synchronized with the real facial motion.




(Figure 1) Block diagram of the proposed 3D facial shape and motion recovery algorithm

## 2. Multi-Modal Fusion Technique

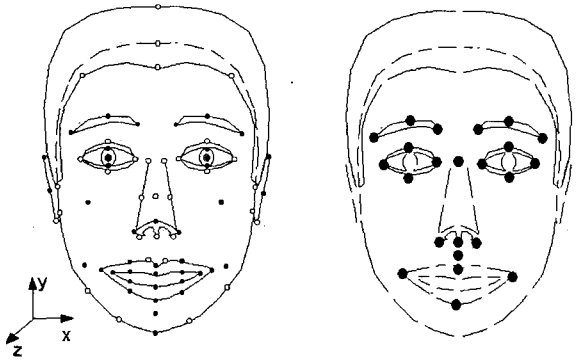
For the robust detection of the facial regions in real environment with various noises, a multi-modal fusion technique is proposed, where range, color and motion information is combined. The summary of proposed algorithm including multi-modal fusion technology is shown in (Figure 1). The information of the range segmented region, facial color transformed region and moving color transformed region are extracted respectively and combined with fuzzy AND operation in all pixels. The resulting image is transformed into a new gray level image whose intensity value is proportion to the probability to be a face. Then pixels of high probability are grouped and segmented as facial objects. It is described in [15] in detail.

## 3. Facial Features Extraction

### 3.1 Facial Features Position

The position of facial features is decided to be satisfied for the MPEG-4 SNHC FDP, which describes practical 3D structure of a face, as shown in (Figure 2). In our work,

23 important features among the MPEG-4 SNHC facial features are selected to be suitable as a stable input values for motion recovery calculation. To extract facial features, generalized skin color distribution (GSCD) and color transform technique are used.




(a) (b)

(Figure 2) Facial feature points (a)proposed by MPEG-4 SNHC and (b)proposed in this paper

### 3.2 Facial moving color information

To exploit the object's motion information, motion detection measurement using UPC (Unmatched Pixel Count) is used. UPC is a block-based method and has simple computational operation [10]. The proposed AWUPC(Aadapted Weighted Unmatched Pixel Count) operation is defined as (1) where $Z(x, y, t)$ is the GSCD transformed image and $U(i, j, t)$ is the UPC motion detected image. The AWUPC operation emphasizes only the region with a motion in the skin-color enhanced region.

$$AWUPC(x, y, t) = Z(x, y, t) \times \sum_{i=x-N}^{i=x+N} \sum_{j=y-N}^{j=y+N} U(i, j, t) \quad (1)$$

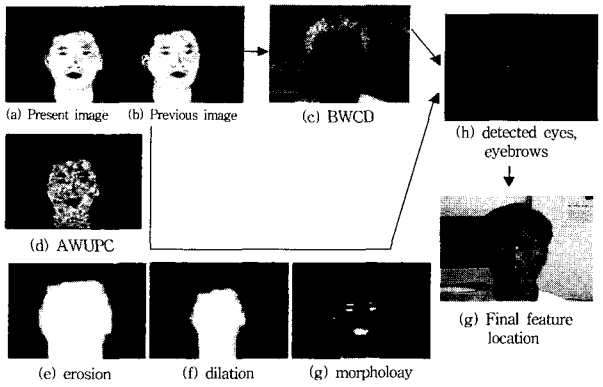
where

$$U(i, j, t) = \begin{cases} 1, & \text{if } |Z(x, y, t) - Z(x, y, t-1)| \le V_{th} \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

Then, to decide the threshold value in (2), a sigmoid function[12] is introduced to induce a adaptive threshold value and is shown as (3)

$$V_{th} = \frac{255}{1 + \exp \dfrac{Z(x, y, t) - \dfrac{255}{2}}{Q}} \quad (3)$$

where $Z(x, y, t)$ is an input pixel value at time t and Q value is a coefficient that decides the slope of threshold curve of sigmoid function. The effectiveness of adaptive threshold value can be described as follows. The input pixel's gray level means the probability of faces. A region that has high probability of faces may be combined with a low threshold value to detect faces very well even when a slight motion. On the contrary, a region with low probability of faces should be combined with a high threshold value to be decided as a face only when it has a large motion.




(Figure 3) A method of detecting eyes and eyebrows by BWCD and top-hat transform

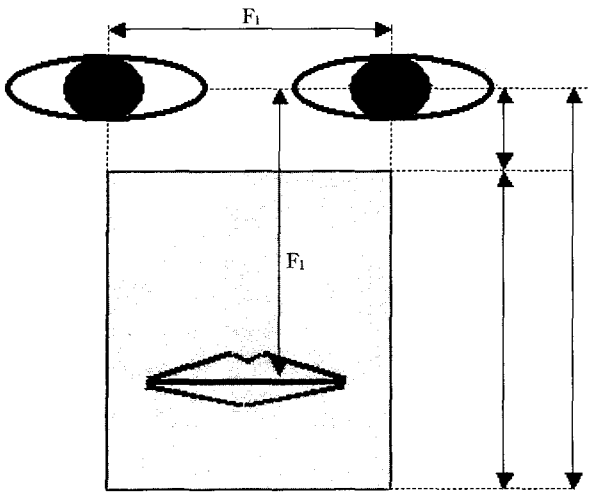
### 3.3 Extraction of Eyes and Eyebrows

The most important facial features are eyes and eyebrows because face has inherent symmetrical shape and color. To detect eyes and eyebrows within a detected facial region, both the BWCD (Black and White Color Distribution) color transform and the AWUPC (Adaptive Weighted Unmatched Pixel Count) moving color detection technique are combined together. Firstly, input image is transformed with BWCD to enhance the eye region as in (Figure 3) (c), which indicates eyes and eyebrows with normalized color space (r, g) = (85, 85). Then, the motion detected image (Figure 3) (d) is transformed by Top-hat operation [14] which enhances only the detail facial components. This Top-hat transformed image is combined with the prepared BWCD enhanced image to detect final eyes and eyebrows (Figure 3) (h). All procedures including Top-hat transform are shown in (Figure 3).

### 3.4 Extraction of Mouth and Nose

If the regions of eyes and eyebrows are decided, the location of mouth can be estimated by general geometrical relation such as (Figure 4). In (Figure 4), the rectangle including the mouth is defined as the candidate window of mouth and nose. F1 indicates the distance between the two ey-

es. In the GSCD transformed images,




(Figure 4) Geometrical relation between eyes and mouth in general faces

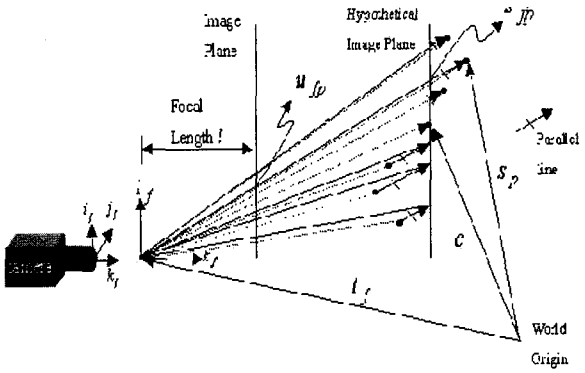
generally mouth region has lower intensity values than the general skin region within the mouth candidate window, so the mouth region can be separated from the skin region with the intensity histogram. Then the image is transformed to binary image to get the exact mouth boundary. (Figure 5) represents 4 points on the mouth boundary line are decided as mouth feature points. Anothor 4 feature points are defined on the nose boundary region. They are extracted using a dominant peak point of pixel's gray level histogram.

(Figure 5) Estimated mouth area and binary image for final decision (a) candidate window for mouth (b) binary image

## 4. Shape and Motion Recovery using a Paraperspective Camera Model

### 4.1 Paraperspective Camera Model

The paraperspective camera model is a linear approximation of the perspective projection by modeling both the scaling effect and the position effect, while retaining the linear properties with a scaling effect, near objects appear larger than the ones with distance, and with a position effect, objects in the periphery of the image are viewed from a different angle than those near the center of projection. As illustrated in figure \ref{fig : para}, the paraperspective projection of an object onto an image involves two steps.




(Figure 6) Paraperspective camera model

Step 1 : An object point, $\vec{s_p}$, is projected along the direction of the line connecting the focal point of the camera and the object's center of mass, $\vec{c}$, onto a hypothetical image plane parallel to the real image plane and passing through the object's center of mass. In frame *f*, each object points $\vec{s_p}$ is projected along the direction of $\vec{c}-\vec{t_f}$ onto the hypothetical image plane where $\vec{t_f}$ is the camera origin with respect to the world origin. If the coordinate unit components of the camera origin are defined as $\vec{i_f}$, $\vec{j_f}$ and $\vec{k_f}$, the result of this projection, $\vec{s_{fp}}$, is given as follows,

$$\vec{s_{fp}} = \vec{s_p} - \frac{(\vec{s_p}\cdot\vec{k_f})-(\vec{c}\cdot\vec{k_f})}{(\vec{c}-\vec{t_f})\cdot\vec{k_f}}(\vec{c}-\vec{t_f}) \qquad (4)$$

Step 2 : The point is then projected onto the real image plane using perspective projection. Since the hypothetical plane is parallel to the real image plane, this is equivalent to simply scaling the point coordinates by the ratio of the camera focal length and the distance between the two planes. Subtracting $\vec{t_f}$ from $\vec{s_{fp}}$, the position of the $\vec{s_{fp}}$ is converted with the camera coordinate system. Then, by scaling the result with the ratio of the camera's focal length $l$ and the depth to the object's center of mass $\vec{z_f}$ results paraperspective projection of $s_{fp}$ onto image plane. By placing the world origin at the object's center of mass, the equation can be simplified without loss of generality and the general paraperspective equation is given as follows,

$$\overrightarrow{u_{fp}} = \frac{1}{z_f}\left\{\left[\vec{i}_f + \frac{\vec{i}_f \cdot \vec{t}_f}{z_f}\vec{k}_f\right]\vec{s}_f - \vec{t}_f \cdot \vec{i}_f\right\} \quad (5)$$

$$\overrightarrow{v_{fp}} = \frac{1}{z_f}\left\{\left[\vec{j}_f + \frac{\vec{j}_f \cdot \vec{t}_f}{z_f}\vec{k}_f\right]\vec{s}_f - \vec{t}_f \cdot \vec{j}_f\right\} \quad (6)$$

where $\overrightarrow{u_{fp}}$ and $\overrightarrow{v_{fp}}$ represents $i$ and $j$ component of the projected point.

Facial feature vectors extracted from above section are converted to $\overrightarrow{u_{fp}}$ and $\overrightarrow{v_{fp}}$ and used to calculate motion information $\overrightarrow{m_f}$, $\overrightarrow{n_f}$ and shape information $\vec{s}_p$. The general paraperspective model equation can be rewritten simply as follows,

$$\overrightarrow{u_{fp}} = \overrightarrow{m_f} \cdot \vec{s}_p + x_f \quad \overrightarrow{v_{fp}} = \overrightarrow{n_f} \cdot \vec{s}_p + y_f \quad (7)$$

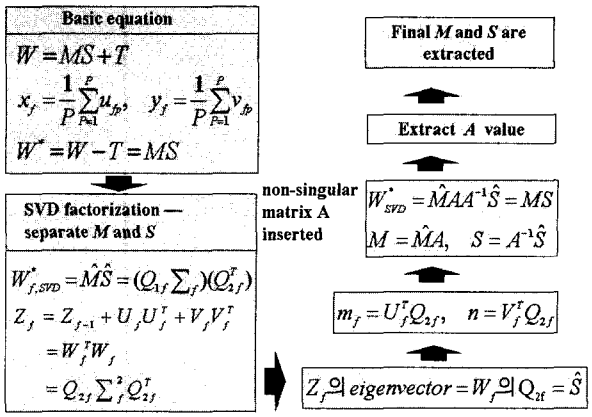
where

$$\overrightarrow{m_f} = \frac{\vec{i}_f - \vec{x}_f \cdot \vec{k}}{z_f}, \quad \overrightarrow{n_f} = \frac{\vec{j}_f - \vec{x}_f \cdot \vec{k}}{z_f} \quad (8)$$

$$z_f = -\vec{t}_f \cdot \vec{k}_f, \quad \vec{x}_f = -\frac{\vec{t}_f \cdot \vec{i}_f}{z_f}, \quad \vec{y}_f = -\frac{\vec{t}_f \cdot \vec{j}_f}{z_f} \quad (9)$$

In equation (7), the term $\overrightarrow{m_f}$ and $\overrightarrow{n_f}$ represent the camera orientation information which means motion information, while $\vec{s}_p$ containing the coordinate of the feature points represents shape information. The $\vec{x}_f$, $\vec{y}_f$ represent the translation information between world origin and camera origin. Notice that the motion information and the shape information can be separated in paraperspective camera model. The 3D feature points, $\vec{s}_p$ , and motion information( $\overrightarrow{m_f}$, $\overrightarrow{n_f}$ ) are calculated by solving equation (7) using Singular Value Decomposition (SVD). Details of the procedure are described in [7].

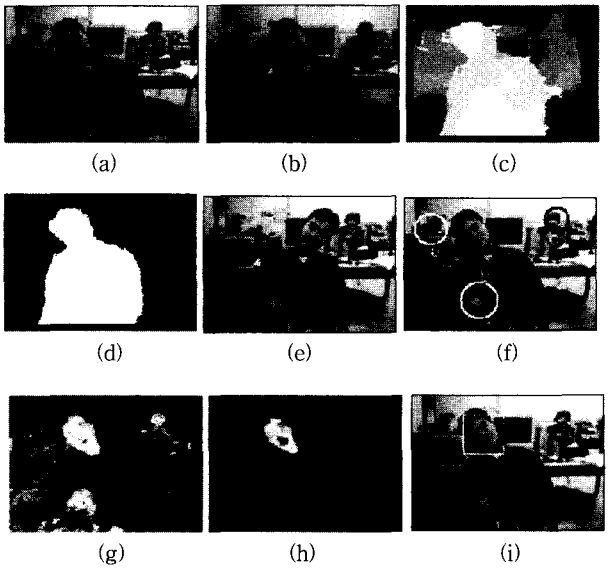
Although the motion and shape recovering algorithm using this paraperspective model shows good result [8], there is large difference between the measured object's depth information and calculated one using a paraperspective camera model which is due to the non-linearity of practical perspective camera model. The method in this paper reduces the depth error lower than the 30% of the measured value. The overal SVD factorization method to recover the shape and motion information is summarized in (Figure 7).




(Figure 7) Summary of the proposed SVD factorization method

## 5. Experimental Results

### 5.1 Facial Object Detection using Multi-Modal Fusion




(Figure 8) face region detection result using range, skin color and motion information : (a), (b) stereo image pairs (t=0) ; (c) MPC disparity map ; (d) range segmented image ; (e)color image sequence (t = -1) (f) inserted noises ; (g)skin color transformed image ; (h) AWUPC transform image ; (i) final face region

(Figure 8) shows the results of multi-modal fusion technique to extract the facial regions. Test images include various skin color noises, complex background and different ranged objects. (Figure 8) (a), (Figure 8) (b) show stereo image pairs at present time (t = 0) for range information and (Figure 8) (e) shows color image sequence at previous time (t = -1) for skin color information. Figure 8(f) shows inserted noises within the (Figure 8) (a) by marking 'a', 'b', 'c'. Mark a represents a object with motion, but without skin

color. Mark b,c represent object with skin color, but without motion. (Figure 8) (c) shows disparity map of (Figure 8) (a), (Figure 8) (b) stereo pairs and (Figure 8) (d) is a range segmented image from (Figure 8) (c). (Figure 8) (g) shows skin color enhanced image using GSCD and (Figure 8) (h) represents the AWUPC result image enhancing only the region having both the motion and skin color,

where skin color noises around human body and skin area having slight motion are removed by using a small motion variable during AWUPC operation. Final face detected regions on input image are shown in (Figure 8) (i).

〈Table 1〉 performance comparison of multi-modal fusion face detection

| Information | No. of test image | No. of successful detection | success ratio |
|---|---|---|---|
| Range | 100 | 46 | 46% |
| Skin color | 100 | 82 | 82% |
| Moving color | 100 | 88 | 88% |
| Multi-modal | 100 | 96 | 96% |

5.2 Performance of the Facial Feature Extraction

The accuracy of automatic facial feature extraction is evaluated by comparing the measured 2D feature location with the extracted values. These results are important in evaluating the performance of recovering 3D shape and motion information because the real face's 3D feature location can not be measured directly. The compared results about error rate are summarized in <Table 2>.

〈Table 2〉 The accuracy of automatic face features extraction, Error ratio = (pixel error/total pixels in each x, y direction

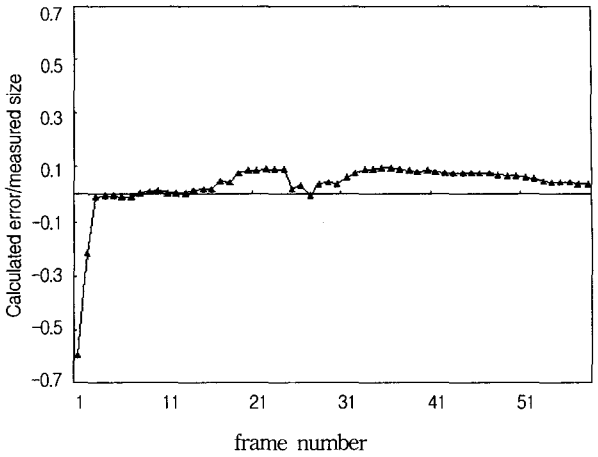
| region | feature number | pixel error (x-dir) | pixel error (y-dir) | error ratio (x-dir) | error ratio (y-dir) |
|---|---|---|---|---|---|
| eyebrow | 6 | 4 | 3 | 1.2% | 1.2% |
| eye | 8 | 5 | 5 | 1.6% | 2% |
| nose | 5 | 7 | 6 | 2.2% | 2.5% |
| mouth | 4 | 4 | 4 | 1.2% | 1.7% |

The success ratio of automatic feature extraction for 2D image sequences are defined as follows when the total number of image frames are 260.

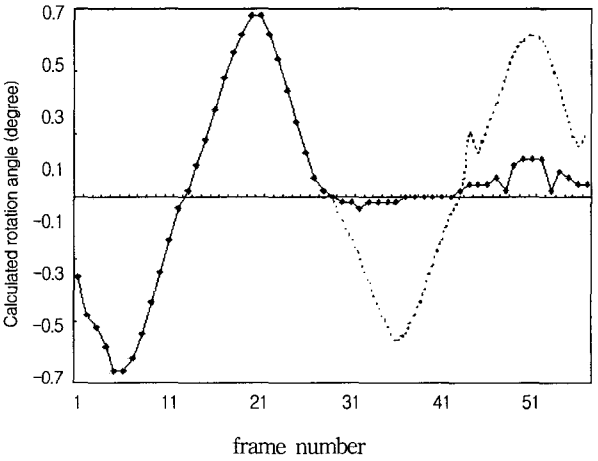
$$success\ \ ratio = \frac{no.of\ \ successful\ \ images}{no.of\ \ test\ \ images} = \frac{252}{260} = 97\% \quad (10)$$

where an image is defined to be successful when its error rate is less than 3% as defined in <Table 2>.

5.3 Accuracy of the Shape and Motion Recovery



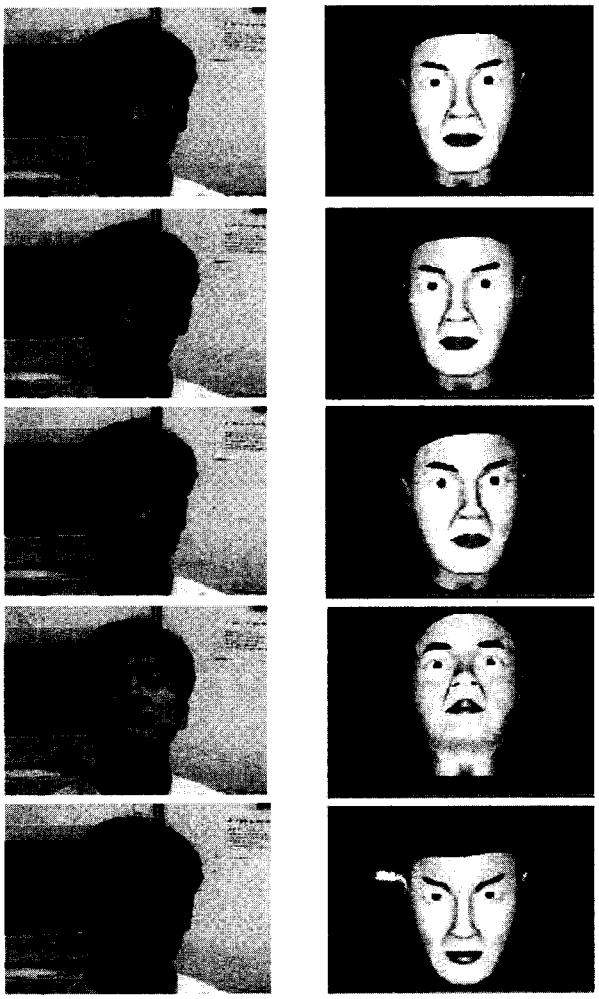

(Figure 9) Comparison of shape recovery for all frames




(Figure 10) Comparison of motion recovery for all frames

To prove the accuracy of the shape and motion recovery, the synthetic pyramid shaped object having 21 feature points along the four edges on the object was used. The synthetic feature points were created by rotating the object, and the test consists of 60 image frames. Rotation of object was presented by three angles, $\alpha, \beta, \gamma$ . The coordinate of center point is defined as (0, 0, 0). The location of left camera is (0, -40, -400) and that of right one is (0, 40, -400). The pyramid object rotates by 15 degree to the left and right around the x-axis and up and down around of the y-axis respectively. The distance between the neighboring features is 20 pixels on the x-axis, and 5 pixels on the y-axis. The result of motion recovery is represented by rotation angle, $\alpha, \beta, \gamma$ as shown in (Figure 10) and shape recovery on the

x, y, z axis for all test frames is shown in (Figure 9). The recovered shape and motion information are transformed automatically to the FAP file of the MPEG-4 to synchronize a generic facial model with a real face. Important results of motion recovery are shown in (Figure 11).



(Figure 11) Motion recovered generic model from real face images ; left column : real input image. right column : motion recovered generic model

## 6. Conclusions

New algorithm that recovers 3D facial features and motion information from 2D stereo image sequences for the MPEG-4 SNHC face model encoding has been proposed. The facial regions are detected using multi-modal fusion technique that combines range, color and motion information. The experiment shows that the success rate of the detection of facial region is over 96% for 100 image frames. The 23 facial features among the MPEG-4 FDP facial features are extracted automatically using morphological processing and moving color transform algorithm(AWUPC). The facial features from 2D image sequences are used to recover the object's 3D shape and global motion sequentially based on the para-perspective camera model SVD factorization method. Finally recovered facial 3D motion and shape information is transformed into the global motion parameters of FAP of the MPEG-4 to synchronize a generic face model with a real face.

## Reference


[1] MPEG-4 System Sub-group, "MPEG-4 System Methodology and Work Plan for Scene Description," ISO/IEC/JTC1/SC29/WG11/N1786, Jul., 1997.

[2] A. Pentland and B. Horowitz. "Recovery of Non-rigid Motion and Structure," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.13, No.7, pp.730-742, 1991.

[3] M. J. Black and Y. Yaccob, "Tracking and Recognizing Rigid and Non-rigid Facial Motion using Local Parametric Model of Image Motion," Proc. Intl Conf. Computer Vision, pp.374-381, 1995.

[4] A. Azarbayejani and A. Pentland, "Recursive Estimation of Motion, Structure and Focal Length," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.7, No.6, pp.562-575, Jun., 1995.

[5] J. Weng, N. Ahuja and T. S. Huang, "Optimal Motion and Structure Estimation," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.15, No.9, Sept., 1993.

[6] T. S. Huang and O. D. Faugeras, "Some Properties of the E-matrix in Two-view Motion Estimation," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.11, No.12, pp.1310-1312, Dec., 1989.

[7] C. J. Poelman and T. Kanade, "A Paraperspective Factorization Method for Shape and Motion Recovery," Technical Report CMU-CS-93-219, Carnegie Mellon University, 1993.

[8] B. O. Jung, "A Sequential Algorithm for 3-D Shape and Motion Recovery from Image Sequences," Thesis for the degree of master, Korea University, Jun., 1997.

[9] Jibe Yang and Alex Waybill, "Tracking Human Faces in Real Time," Technical Report CMU-CS-95-210, Carnage Melon University, 1995.

[10] H. Gharavi and Mike Mills "Blockmatching Motion Estimation Algorithm-New Results," IEEE Trans. Circuits and System, No.5, Vol.37, 1990.

[11] S. H. Kim, H. G. Kim and K. H. Tchah, "Object-oriented Face Detection using Colour Transformation and Range Segmentation," IEE Electronics Letters, 14th, Vol.34, No.10,

pp.979-980, May, 1998.

[12] D. reisfild, "Detection and Interest Points using Symmetry," Proc. Intl Conf. Computer Vision, Vol.E81-D, pp.62-65, Dec., 1990.

[13] MPEG-4 SNHC Group, "Face and Body Definition and Animation Parameter," ISO/IEC JTC1/SC29/WG11 N2202, March, 1998.

[14] R. C. Gonzalez and R. E. Woods, "Digital Image Processing," Addison-Wesley, pp.225-238, 1992.

[15] S. H. Kim and H. G. Kim. "Face Detection using Multi-Modal Information," Proc. Intl Conf. Face and Gesture Recognition, France, March, 2000.

**김 상 훈**

e-mail : kimsh@hnu.hankyong.ac.kr

1987년 고려대학교 전자공학과 학사

1989년 고려대학교 대학원 전자공학과 석사

1999년 고려대학교 대학원 전자공학과 박사

2001년 6월 UWA 방문연구원(Australia)

1989년~1994년 LG반도체 연구원

1999년~2001년 KIST 위촉연구원

1999년~현재 국립한경대학교 제어계측공학과 조교수

관심분야 : 3D 영상처리, face detection, real-time object tracking