3—9 Simultaneous Estimation of Head Pose and Shape by Hierarchical Control System

Junji SATAKE Takeshi SHAKUNAGA*
Department of Information Technology, Okayama University

Abstract

The present paper proposes a simultaneous estimation of both human head pose and shape. When head pose is estimated for a generic face model, a considerable estimation error is generated by the shape difference between the model and the target face. In the proposed framework, a specific face model is gradually created for the target face by deforming the generic model based on an error analysis. These tasks are controlled by a hierarchical attention control system. Stable and efficient estimation is realized by switching the processes in a hierarchical control system. Experimental results show that the proposed system can simultaneously estimate head pose and shape for several persons.

1 Introduction

Pose estimation and shape estimation are the most important problems in computer vision. The present paper proposes a simultaneous estimation of head pose and shape. Several studies [1, 2] have proposed methods for head pose estimation. However, most head pose estimation methods are based on a generic face model, resulting in considerable error due to the shape difference between the model and the target face.

In order to estimate head pose, the shape of the target face must be obtained. Hence, process switching between pose and shape estimation is necessary. The present paper proposes a simultaneous estimation of head pose and shape by process control using a hierarchical control system. Therefore, the correct head pose can be obtained for any person, as well as for the specific model of the target face.

2 Simultaneous estimation of head pose and shape

2.1 Complementation of pose and shape estimation

In order to estimate head pose, switching between estimation of head pose and shape is necessary. In addition, the feature detection processes should be controlled according the direction of the face. The estimation of correct head pose and shape is realized by these task controls.

2.2 Control of pose and shape estimation

The present paper proposes task control based on a hierarchical control system [3]. This section describes process switching using a hierarchical control system for stable and efficient estimation.

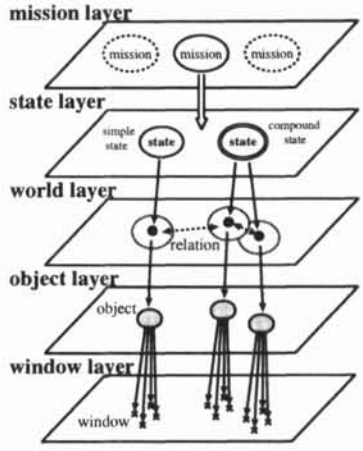

Figure 1: Hierarchical attention control system.

2.2.1 Hierarchical attention control system

Tasks are controlled by a hierarchical attention control system [3], as shown in Fig.1, which consists of five layers. The roles of each layer are summarized as follows:

1) mission layer: This layer controls the mission of the entire system. Missions are switched between according to the mission transition diagram in this layer. The lower four layers are controlled according to the mission.

2) state layer: This layer controls the state of each object. The state is decided according to the state transition diagram by analyzing the result of image processings in the lower layers. In the present paper, facial features (e.g. left eye, right eye, nose, mouth) are considered as objects.

3) world layer: This layer deals with information of relationships among objects. Both head pose and shape are estimated based on the relationships between feature positions on the image.

4) object layer: This layer integrates the information on each object. The stable and efficient process is realized by switching processes or data according to the state.

5) window layer: This layer controls each attention window which is a small region in the image. The image processing is accomplished in each window. The result of image processing is sent to the object layer.

Stable and efficient process control is realized using the hierarchical attention control system. In the bottom-up process, the state of each object is decided

*Address: Tsushima-naka 3-1-1, Okayama, 700-8530, Japan. E-mail: {satake,shaku}@chino.it.okayama-u.ac.jp

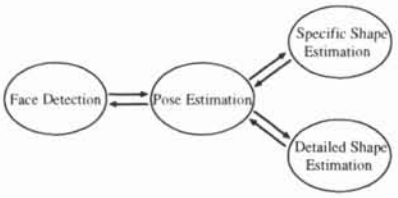

Figure 2: Mission transition diagram.

by analyzing the result of image processings. In the top-down process, the image processings are switched and the attention is controlled according to the state. The task of the entire system is controlled in the mission layer, and the image processings are switched according to the mission in the world layer. On the other hand, the state of each object is decided in the state layer, and the processing for the object is switched according to the state in the object layer.

2.2.2 Mission control of entire system (mission layer)

The mission control of the entire system is realized by switching the state in the mission layer, and the state of the entire system is called the mission. In the present paper, four missions, namely a) face detection, b) pose estimation, c) specific shape estimation, and d) detailed shape estimation, are defined. The tasks on each mission are summarized as follows:

a) Face detection A face is detected.

b) Pose estimation Facial features are detected, and head pose is estimated.

c) Specific shape estimation The specific face model of the target face is estimated.

d) Detailed shape estimation The detailed face model is estimated.

The specific face model is specific shape information of the target face which is invariable with respect to time. On the other hand, the detailed face model is detailed shape information (e.g. mouth shape) which varies with facial expression over time.

The mission is switched depending on the process situation, which is managed in the world layer. The mission transition diagram is shown in Fig.2. The mission control of the entire system is summarized as follows:

- If a head is not detected at previous time, a) face detection is continued.
- If a head is detected at previous time, b) pose estimation is performed. In initial pose estimation, the generic face model is used.
- When the number of frames having the exact detected feature positions exceeds a threshold, c) specific shape estimation is performed.
- Then, b) pose estimation is performed using the obtained specific face model. The system repeats the switching of pose estimation and specific shape estimation.
- If a head pose is estimated precisely using the specific face model, d) detailed shape estimation is performed.

2.2.3 Feature detection (object layer)

Facial features are considered as objects. Because facial features vary depending on head pose, detection processes are switched according to head pose. In this system, three templates of each feature are switched according to head direction (i.e. front, left, right).

2.2.4 Control of feature detection (state layer)

Because head pose cannot be estimated from incorrect feature information, stable feature detection is necessary. Therefore, feature detection processes are switched according to head direction. This switching is controlled according to the state in the state layer.

2.2.5 Use of face structure information (world layer)

In the world layer, the processes which use face structure information are performed. Head pose is quickly estimated from detected feature points via an extended Kalman filter. Whereas head shape is estimated by deformation of the generic model based on an error analysis. The tangible processes are described in section 3.

The process of each object in the object layer is switched according to the state in the state layer, whereas the process in the world layer is switched according to the mission in the mission layer.

3 Estimation of head pose and shape

3.1 Face detection

This section shows the processes in the world layer for each mission.

Judgment as to whether a face exists on the image is performed. Face detection involves the following information: background subtraction, skin color, and an ellipsoidal shape.

3.2 Pose estimation via EKF

This section shows head pose estimation via an extended Kalman filter. The state transition equation and the measurement equation are defined, and a method of head pose estimation is described.

3.2.1 State transition equation

As head pose parameters, the 3-D position (X_0, Y_0, Z_0) and directions ψ, θ, ϕ are defined as shown in Fig.3, and the state variable $\boldsymbol{\xi}_k$ which denotes the head pose at instant k is defined as

$$\boldsymbol{\xi}_k = \left[\begin{array}{cccccc} X_0 & Y_0 & Z_0 & \psi & \theta & \phi \end{array} \right]^T,$$

where (X_0, Y_0, Z_0) is in the camera-centered coordinate system, and ψ, θ, ϕ respectively denote rotations on the X, Y, Z coordinate system.

Then, $\dot{\boldsymbol{\xi}}_k$ denotes a differential of $\boldsymbol{\xi}_k$, and $\ddot{\boldsymbol{\xi}}_k$ denotes a differential of $\dot{\boldsymbol{\xi}}_k$. If Δt is small, a head movement can be considered as a uniformly accelerated motion, and the state transition equation is defined as

$$\left[\begin{array}{c} \boldsymbol{\xi}_{k+1} \\ \dot{\boldsymbol{\xi}}_{k+1} \\ \ddot{\boldsymbol{\xi}}_{k+1} \end{array} \right] = \boldsymbol{F} \left[\begin{array}{c} \boldsymbol{\xi}_k \\ \dot{\boldsymbol{\xi}}_k \\ \ddot{\boldsymbol{\xi}}_k \end{array} \right] + \boldsymbol{u}_k, \qquad (1)$$

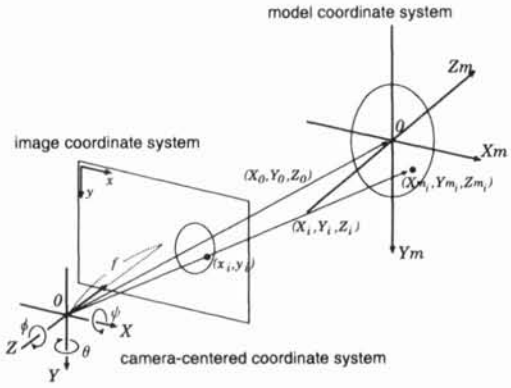

Figure 3: Coordinate systems.

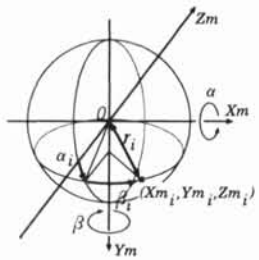

Figure 4: Polar coordinates of the face model.

where

$$\boldsymbol{F} = \begin{bmatrix} \boldsymbol{I} & \Delta t\boldsymbol{I} & \frac{\Delta t^2}{2}\boldsymbol{I} \\ \boldsymbol{0} & \boldsymbol{I} & \Delta t\boldsymbol{I} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I} \end{bmatrix},$$

$\boldsymbol{I}$ is a 6 × 6 identity matrix, and $\boldsymbol{u}_k$ is a noise. By considering $\boldsymbol{u}_k$, a tracking which copes with a change in acceleration is realized.

3.2.2 Measurement equation

The i-th feature point on the generic model is denoted as (X_i, Y_i, Z_i) in the camera-centered coordinate system, and as $(X_{m_i}, Y_{m_i}, Z_{m_i})$ in the model coordinate system. Then, the following equation is obtained.

$$\begin{bmatrix} X_i \\ Y_i \\ Z_i \end{bmatrix} = \begin{bmatrix} X_0 \\ Y_0 \\ Z_0 \end{bmatrix} + \boldsymbol{R} \begin{bmatrix} X_{m_i} \\ Y_{m_i} \\ Z_{m_i} \end{bmatrix},$$

where $\boldsymbol{R} = \boldsymbol{R}_Z(\phi)\boldsymbol{R}_X(\psi)\boldsymbol{R}_Y(\theta)$, and $\boldsymbol{R}_X(\psi)$, $\boldsymbol{R}_Y(\theta)$, $\boldsymbol{R}_Z(\phi)$ respectively denote rotations ψ, θ, ϕ on the X, Y, and Z axes.

The model coordinate system is expressed in the polar coordinate system (α_i, β_i, r_i), as shown in Fig.4 for simplification, where α_i and β_i are the rotation angles on the X_m and Y_m axes, respectively, and $r_i = ({X_{m_i}}^2 + {Y_{m_i}}^2 + {Z_{m_i}}^2)^{\frac{1}{2}}$. In the weak perspective projection, the projection coordinate (x_i, y_i) of the i-th feature point (α_i, β_i, r_i) is given as

$$\begin{aligned} x_i &= \frac{f}{Z_0}X_0 - \frac{fr_i}{Z_0}\{s_\phi c_\psi s_{\alpha_i} \\ &\quad + c_\phi s_{(\theta+\beta_i)} c_{\alpha_i} + s_\phi s_\psi c_{(\theta+\beta_i)} c_{\alpha_i}\}, \qquad (2) \\ y_i &= \frac{f}{Z_0}Y_0 + \frac{fr_i}{Z_0}\{c_\phi c_\psi s_{\alpha_i} \\ &\quad - s_\phi s_{(\theta+\beta_i)} c_{\alpha_i} + c_\phi s_\psi c_{(\theta+\beta_i)} c_{\alpha_i}\}, \qquad (3) \end{aligned}$$

where f is the focal length, $\sin\alpha = s_\alpha$, and $\cos\alpha = c_\alpha$.

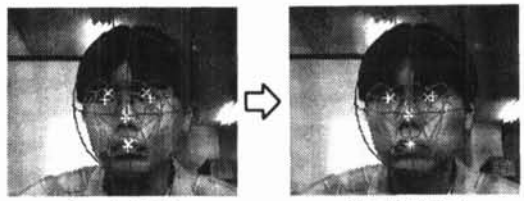

Figure 5: Estimation of head shape.

In addition, by lining up n feature coordinates $(x_i^*, y_i^*)(i = 1, 2, \cdots, n)$ observed on the image at instant k, the $2n$-D measurement variable $\boldsymbol{\eta}_k$ is created as

$$\boldsymbol{\eta}_k = [x_1^*\ y_1^*\ x_2^*\ y_2^*\ \cdots\ x_n^*\ y_n^*]^T. \qquad (4)$$

From equations (2)–(4), the measurement equation is given as

$$\boldsymbol{\eta}_k = \boldsymbol{h}_k(\boldsymbol{\xi}_k) + \boldsymbol{v}_k, \qquad (5)$$

where

$$\boldsymbol{h}_k(\boldsymbol{\xi}_k) = [x_1\ y_1\ x_2\ y_2\ \cdots\ x_n\ y_n]^T,$$

and $\boldsymbol{v}_k$ denotes the observation noise.

3.2.3 Estimation of pose parameter

The feature points are detected on the image, and the measurement variable $\boldsymbol{\eta}_k$ is created. Then, the state transition equation (1) and the measurement equation (5) are quickly solved via an extended Kalman filter [4]. The obtained state variable $\boldsymbol{\xi}_k$ denotes the head pose at instant k.

3.2.4 Error detection of feature

If feature positions have noises, the head pose cannot be estimated precisely. Therefore, feature errors must be detected in the world layer. This error detection is performed by judgment of the position relationship between the features using face structure information. Feature positions with error are replaced with prediction values using a Kalman filter and are not used in shape estimation.

3.3 Specific shape estimation

This section shows the head shape estimation for a target face. If the generic face model differs from the shape of the target face, an observation error is generated between the observed feature coordinate (x_i^*, y_i^*) and the projected feature coordinate $(\hat{x}_i, \hat{y}_i)$ from the estimated head pose $\hat{\boldsymbol{\xi}}_k$. In this section, estimation of the specific face model, which is gradually created by deforming the generic model based on an error analysis, as shown in Fig.5, is described.

Using the coordinates (α_i, β_i, r_i) of feature points in the generic face model, the coordinates of the specific face model are defined as $(\alpha_i + \Delta\alpha_i,\ \beta_i + \Delta\beta_i,\ r_i + \Delta r_i)$. The observed coordinates (x_i^*, y_i^*) of feature points are considered as projections of this specific face model. From equations (2) and (3), using the error between the observed coordinates and the projected coordinates $(\hat{x}_i, \hat{y}_i)$ of the feature points in the generic face

model, the correction values $\Delta\alpha_i$, $\Delta\beta_i$, and Δr_i are calculated. If the quadratic and cubic terms for Δ are approximately equal to 0, the following linear exists:

$$\begin{bmatrix} x_i^* - \hat{x}_i \\ y_i^* - \hat{y}_i \end{bmatrix} = \left. \begin{bmatrix} x_i \\ y_i \end{bmatrix} \right|_{\substack{\alpha_i + \Delta\alpha_i \\ \beta_i + \Delta\beta_i \\ r_i + \Delta r_i}} - \left. \begin{bmatrix} x_i \\ y_i \end{bmatrix} \right|_{\substack{\alpha_i \\ \beta_i \\ r_i}} \approx \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{bmatrix} \Delta\alpha_i \\ \Delta\beta_i \\ \Delta r_i \end{bmatrix}, \quad (6)$$

where

$$a_{11} = -\frac{f r_i}{Z_0}\{c_\phi c_{(\theta+\beta_i)} c_{\alpha_i} - s_\phi s_\psi s_{(\theta+\beta_i)} c_{\alpha_i}\}, \quad etc.$$

Equation (6) exists at each frame, and for each feature point. From multiple frame information, in which head pose changes, and for which feature points are exactly detected, simultaneous equations are obtained. The correction values, $\Delta\alpha_i, \Delta\beta_i, \Delta r_i$ for each feature point are calculated via a least-squares method.

3.4 Detailed shape estimation

This section shows a detailed shape estimation. In creating the specific face model, although the positions of features (e.g. eye, nose) can be roughly estimated, detailed shapes (e.g. contours of the mouth) have noises. These noises are generated from the personal shape differences or deformation by facial expression. Therefore, a more detailed shape is estimated in this section.

In the model coordinate system, an edge is detected radially based on the centered 3-D coordinate $(X_{m_i}, Y_{m_i}, Z_{m_i})$ of the i-th feature point. Therefore, the outlines of facial features are estimated. By edge detection for each direction in which direction vectors are projected on the image, the 3-D coordinates of outline are estimated.

4 Experimental results

Figure 6 shows the experimental results of pose and shape estimations. These results show that the shape and pose are simultaneously estimated each time. After feature detection, 0.2 ms per frame is required for pose estimation and 0.5 ms is required for shape estimation using a PC containing an Athlon 1.0-GHz processor. Figure 7 shows the RMS errors for the eyes on an image sequence. In this figure, pose estimation was started at the 15th frame. The positional error, which is caused by shape difference between the generic model and the target face during the 19th–34th frames, is approximately 8 pixels. After specific shape estimation at the 35th frame, the error decreases to approximately 2 pixels, which seems to be caused by observation noise. This result also shows that the specific face model is correctly generated from the generic model.

5 Conclusions

The present paper described a simultaneous estimation of head pose and shape by a hierarchical control system. Using the proposed estimation, a specific face model and a correct pose can be obtained for any person. Stable and efficient process control is realized using a hierarchical control system. Future research will investigate object recognition based on a generic object model.

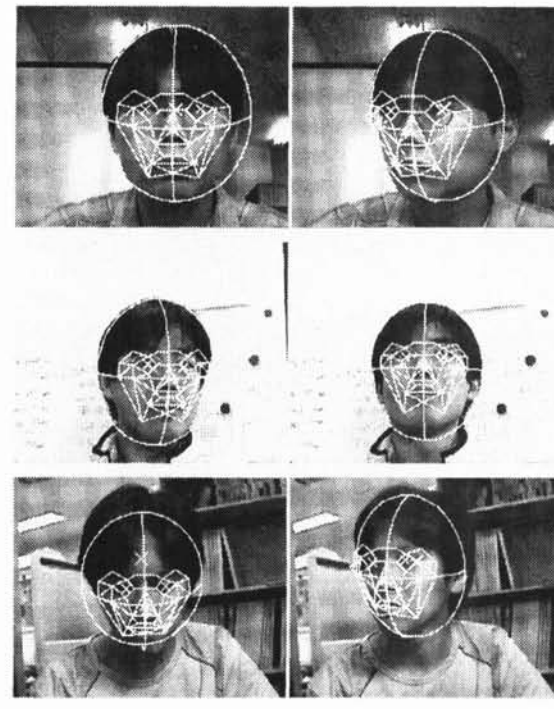

Figure 6: Results of pose and shape estimations.

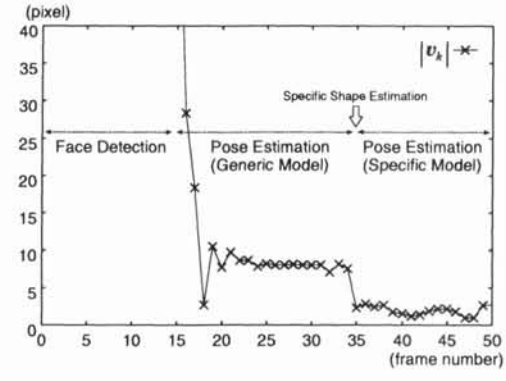

Figure 7: Positional errors with eyes on image sequence.

Acknowledgment

This work has been supported in part by Japan Science and Technology Corporation under Ikeuchi CREST project.

References

[1] C. S. Wiles, A. Maki and N. Mastuda, "Hyperpatches for 3D Model Acquision and Tracking," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.23, No.12, pp.1391–1403, Dec. 2001.

[2] S. Basu, I. Essa and A. Pentland, "Motion Regulation for Model-based Head Tracking," Proc. Int'l Conf. Pattern Recognition, Aug. 1996.

[3] J. Satake and T. Shakunaga, "Human Tracking Based on Hierarchical Attention Control", Proc. 2000 IAPR Workshop on Machine Vision Applications (MVA2000), pp.517-520, Nov. 2000.

[4] Z. Zhang and O. Faugeras, 3D Dynamic Scene Analysis, Springer-Verlag, 1992.