MAXIMUM-LIKELIHOOD MOTION ESTIMATION OF A HUMAN FACE

Geovanni Martinez
Universidad de Costa Rica, Escuela de Ingenieria Electrica
2060 San Jose, Costa Rica, Central America
gmartin@pacuare.eie.ucr.ac.cr

ABSTRACT
An algorithm for estimating the three-dimensional motion of a human face from a monocular image sequence is investigated. For motion estimation, the shape of a human face is described by a three-dimensional rigid triangular mesh and its motion by six parameters: one three-dimensional translation vector and three rotation angles. The motion parameters are estimated by maximizing the conditional probability of the frame to frame intensity differences at observation points. The conditional probability is a function of the motion parameters, the frame to frame intensity differences and the covariance matrix of the intensity error at the observation points. The intensity error is supposed to be the result of the camera noise and the position error attributed to the shape estimation errors and the motion estimation errors occurred by the motion analysis of previous frames. The algorithm was applied to different real image sequences depicting a moving human face with very encouraging results.

1. INTRODUCTION
In the last decades human motion estimation from a monocular image sequence has become a very important research area particularly for video compression [1, 2, 3, 4] and teleoperation control of remotely operated humanoid robots [5]. The existing algorithms for human motion estimation can be roughly divided into two groups. The first one extracts and tracks image features (e.g. edge points) and then determines the motion from these correspondences [6, 7, 8]. The second one estimates the motion by minimizing the frame to frame intensity differences at observation points [9, 10, 3, 11]. In [12] both feature points correspondences and intensity differences are taken into account for motion estimation. In these contribution, the motion is estimated by maximizing the conditional probability of the frame to frame intensity differences at observation points [5, 13, 2]. The Maximum-Likelihood motion estimation algorithm was first proposed in [13, 2] for estimating the motion of the head and shoulders of a person for Object-Based Analysis Synthesis Coding and then used in [5] for estimating the motion of the right arm of a person for teleoperation space applications. In this contribution, we use the Maximum-Likelihood motion estimation algorithm for estimating the motion of a person’s face only.

The Maximum-Likelihood motion estimation algorithm is applied to different real image sequences depicting a moving face to assess its accuracy, limitations and advantages. This paper is organized as follows. In section 2, the Maximum-Likelihood motion estimation algorithm is described. In section 3, some experimental results with real image sequences are presented. In section 4, the conclusions are given.

2. MAXIMUM-LIKELIHOOD MOTION ESTIMATION
For Maximum-Likelihood motion estimation, the shape of a face is considered to be rigid and described by the triangular mesh proposed in [14]. At the beginning of the image sequence, the shape and pose of the triangular mesh is adapted to the content of the first image of the image sequence by applying the automatic face model adaptation algorithm proposed in [15]. The color of a face is described by the intensity and chrominance values on the face surface and is defined by projecting a real image back onto the triangular mesh. The illumination is supposed to be diffuse and perspective projection with focal distance F is used. The three-dimensional motion of a human face from time t_k to t_{k+1} is described by six parameters B_{k\rightarrow k+1} = (T_X, T_Y, T_Z, R_X, R_Y, R_Z)^T: one three-dimensional translation vector T_{k\rightarrow k+1} = (T_X, T_Y, T_Z)^T and three rotation angles R_X, R_Y, R_Z. An arbitrary point H_k = (H_X, H_Y, H_Z)^T on the surface of the face moves to its new position H_{k+1} = (H'_{X}, H'_{Y}, H'_{Z})^T according to the following motion equation:

\[ H_{k+1} = R_{k\rightarrow k+1} \cdot (H_k - G_k) + G_k + T_{k\rightarrow k+1}, \]  

where R_{k\rightarrow k+1} is the rotation matrix defined by the three rotation angles and G_k is the origin of the local coordinate

Thanks to Dr.-Ing. Markus Kampmann for facilitating the stp files of the face’s triangular mesh already adapted to the first image of each one of the real image sequences used for the experimental results.
system of the human face. In following an algorithm for estimating the motion parameters $B_{k-k+1}$ is described.

For estimating the motion parameters $B_{k-k+1}$ of observation points $W_{k,j}$, $0 \ldots J-1$, $J > 6$, is evaluated. This set of observation points is created only ones after face model adaptation. Each observation point $W_{k,j} = (H_{k,j}^{x}, g_{k,j}^{x}, I(j))$ lies on the surface of the triangular mesh and is described by its position $H_{k,j}^{x} = (H_{k,j}^{x}, H_{k,j}^{y}, H_{k,j}^{z})^T$ with respect to the world coordinate system, the intensity value $I(j)$ at this position and the linear intensity gradients $g_{k,j}^{x} = (g_{k,j}^{x}, g_{k,j}^{y})$. The intensity values and the linear intensity gradients are taken from the same image used for face model adaptation. In order to reduce the influence of the camera noise and to increase the accuracy of the estimates only points with high linear intensity gradients ($|g| > \delta_{l}$) are part of the set of observation points. The linear intensity gradients are computed using the Sobel operator. The perspective projection of an observation point $W_{k,j}$ into the image plane is defined as $h_{k,j}^{x} = \begin{pmatrix} F \cdot H_{k,j}^{x} / H_{Z} \cdot F \cdot H_{Y} / H_{Z} \end{pmatrix}^T = (h_{k,j}^{x}, h_{k,j}^{y})^T$. Assuming that the shape and pose of the triangular mesh of the face correspond to those of the real face at time $t_k$, the frame to frame intensity difference $fd(h_{k,j}^{x})$ at the observation point $h_{k,j}^{x}$ is approximated as follows:

$$fd(h_{k,j}^{x}) = I_{k+1}(h_{k,j}^{x}) - I_{k}(h_{k,j}^{x}) \approx I_{k+1}(h_{k,j}^{x}) - I(j), \quad (2)$$

where $I_{k+1}(h_{k,j}^{x})$ and $I_{k}(h_{k,j}^{x})$ represent the intensity values of the intensity images $I_{k+1}$ and $I_{k}$ at position $h_{k,j}^{x}$, respectively. Because in general $h_{k,j}^{x}$ lies outside of the image raster, the intensity value $I_{k+1}(h_{k,j}^{x})$ is computed by linear interpolation of the intensity values of the nearest four pixels in the intensity image $I_{k+1}$. The frame to frame intensity difference at the $J$ observation points is represented as follows:

$$FD = (fd(h_{k,j}^{x-1}), fd(h_{k,j}^{x-2}), \ldots, fd(h_{k,j}^{x}))^T. \quad (3)$$

Assuming that the frame to frame intensity difference $fd(h_{k,j}^{x})$ at the observation point $h_{k,j}^{x}$ is due to the object motion only and modelling the intensity signal around the observation point $h_{k,j}^{x}$ by a second order Taylor expansion, the following relationship is written between the frame to frame intensity difference and the positions $h_{k,j}^{x}$ and $h_{k,j}^{x+1}$:

$$fd(h_{k,j}^{x}) = -\frac{\partial(I(j))}{\partial h_{k,j}^{x}} \cdot (h_{k,j}^{x+1} - h_{k,j}^{x}) + \Delta I(j), \quad (4)$$

where $\frac{\partial(I(j))}{\partial h_{k,j}^{x}} = 1/2 \cdot (g_{k,j}^{x} + g_{x+1}(h_{k,j}^{x+1})) = \begin{pmatrix} g_{k,j}^{x} \ g_{k,j}^{y} \end{pmatrix}$, $g_{k+1}(h_{k,j}^{x+1})$ are the linear intensity gradients computed convolving the intensity image $I_{k+1}$ at position $h_{k,j}^{x}$ with the Sobel operator, and $\Delta I(j)$ is the intensity error due to the camera noise and the observation point position error $\Delta H_{k,j}^{(j)}$.

The position error $\Delta H_{k,j}^{(j)}$ is supposed to be the result of the shape estimation errors and motion estimation errors occurred by the motion analysis of previous frames. The position error $\Delta H_{k,j}^{(j)}$ is modelled by a zero mean Gaussian stochastic process with the following covariance matrix:

$$E[\Delta H_{k,j}^{(j)} \cdot \Delta H_{k,j}^{(j)^T}] = \begin{bmatrix} \sigma_{x}^{2} & 0 & 0 \\ 0 & \sigma_{y}^{2} & 0 \\ 0 & 0 & \sigma_{z}^{2} \end{bmatrix}. \quad (5)$$

Modelling the camera and the intensity signal with first order Taylor expansions at position $h_{k,j}^{x}$ and assuming adding camera noise with variance $\sigma_{n}^{2}$, the variance of the intensity error at position $h_{k,j}^{x}$ is written as follows [2]:

$$\sigma_{\Delta I(j)}^{2} = \frac{\sigma_{x}^{2}}{H_{x}^{2}} ((g_{x}^{(j)})^{2} + (g_{y}^{(j)})^{2} + \frac{F^{2}}{\sigma_{z}^{2}} (g_{x}^{(j)} g_{y}^{(j)} + g_{y}^{(j)} g_{x}^{(j)}) + \sigma_{n}^{2}. \quad (6)$$

Assuming that the intensity errors $V = (\Delta I^{(j-1)}, \Delta I^{(j-2)}, \ldots, \Delta I^{(0)})^T$ at the observation points are statistically independent, the corresponding covariance matrix $U = E[V \cdot V^T]$ is written as follows:

$$U = \begin{bmatrix} \sigma_{\Delta I^{(j-1)}}^{2} & 0 & \ldots & 0 \\ 0 & \sigma_{\Delta I^{(j-2)}}^{2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_{\Delta I^{(0)}}^{2} \end{bmatrix}. \quad (7)$$

Replacing $h_{k,j}^{x}$ and $h_{k,j}^{x+1}$, in Eq. 3 with their corresponding coordinates at the world coordinate system and assuming small rotation angles, the conditional probability of the frame to frame intensity differences is written as follows [2]:

$$p(FD|B_{k-k+1}) = e^{-\frac{1}{2} (FD - [O] B_{k-k+1})^T U^{-1} (FD - [O] B_{k-k+1})} / \sqrt{(2\pi)^J |U|}, \quad (8)$$

where $O$ is the system matrix defined as follows:

$$O = ([0^{(j-1)T}, 0^{(j-2)T}, \ldots, 0^{(0)T})^T, \quad (9)$$

and $O^{(j)T} =$

$$\begin{bmatrix} -\frac{\partial(I(j))}{\partial h_{k,j}^{x}} F/H_{x}^{2} \\ -\frac{\partial(I(j))}{\partial h_{k,j}^{y}} F/H_{y}^{2} \\ -\frac{\partial(I(j))}{\partial h_{k,j}^{z}} F/H_{z}^{2} \\ +[H_{x}^{2} \frac{\partial^{2}(I(j))}{\partial h_{k,j}^{x} \partial h_{k,j}^{x}} F/H_{x}^{2} \\ +[H_{y}^{2} \frac{\partial^{2}(I(j))}{\partial h_{k,j}^{y} \partial h_{k,j}^{y}} F/H_{y}^{2} \\ +[H_{z}^{2} \frac{\partial^{2}(I(j))}{\partial h_{k,j}^{z} \partial h_{k,j}^{z}} F/H_{z}^{2} \\ +[H_{x}^{2} \frac{\partial^{2}(I(j))}{\partial h_{k,j}^{y} \partial h_{k,j}^{y}} F/H_{x}^{2} \\ +[H_{y}^{2} \frac{\partial^{2}(I(j))}{\partial h_{k,j}^{y} \partial h_{k,j}^{y}} F/H_{y}^{2} \\ +[H_{z}^{2} \frac{\partial^{2}(I(j))}{\partial h_{k,j}^{z} \partial h_{k,j}^{z}} F/H_{z}^{2} \\ +[H_{x}^{2} \frac{\partial^{2}(I(j))}{\partial h_{k,j}^{y} \partial h_{k,j}^{y}} F/H_{x}^{2} \\ +[H_{y}^{2} \frac{\partial^{2}(I(j))}{\partial h_{k,j}^{y} \partial h_{k,j}^{y}} F/H_{y}^{2} \\ +[H_{z}^{2} \frac{\partial^{2}(I(j))}{\partial h_{k,j}^{z} \partial h_{k,j}^{z}} F/H_{z}^{2} \end{bmatrix}.$$
The Maximum-Likelihood motion estimates \( \hat{B}_{k \rightarrow k+1} \) are determined by maximizing the conditional probability of the frame to frame intensity difference at the \( J \) observation points:

\[
p(FD/\hat{B}_{k \rightarrow k+1}) \geq p(FD/B_{k \rightarrow k+1}) \quad \forall \, B_{k \rightarrow k+1}.
\]

(10)

For convenience Eq. 10 is written as follows [2]:

\[
\frac{\partial}{\partial B_{k \rightarrow k+1}} \{ \ln p(FD/B_{k \rightarrow k+1}) \} |_{B_{k \rightarrow k+1} = \hat{B}_{k \rightarrow k+1}} = 0.
\]

(11)

Thus, the Maximum-Likelihood motion estimates are given by:

\[
\hat{B}_{k \rightarrow k+1} = (O^T U^{-1} O)^{-1} O^T U^{-1} FD.
\]

(12)

In order to improve the reliability and accuracy of the estimates, the algorithm is applied iteratively. The resulting estimates \( 1^\text{st} \hat{B} \) for each iteration \( i \) are used to update the motion estimates \( \hat{B} \) found by previous iterations. After each iteration, the triangular mesh and its observation points are moved using the current estimates. Due to the motion compensation, the frame to frame intensity differences at the observation points decreases. The iteration ends when after two consecutive iterations the mean square frame to frame intensity difference at the observation points does not increase significantly.

3. EXPERIMENTAL RESULTS

We have implemented the Maximum-Likelihood motion estimation algorithm and performed a number of experiments on different real image sequences in CIF format at a frame rate of 10 Hz to assess its accuracy, limitations and advantages for estimating the motion of a moving human face. Due to the lack of space, we just present the experimental results obtained from two different image sequences: Claire (150 frames) and Miss America (50 Frames).

In this experiment, \( \delta_1, \sigma_2^2, \sigma_X^2, \sigma_Y^2 \) and \( \sigma_Z^2 \) were set to 5, 1, \( 10^{-10} \), \( 10^{-10} \) and \( 10^{-10} \), respectively. All these values were experimentally determined. The experiment was performed on an AMD Athlon(tm) MP (1.2GHz) workstation with 2.0 GB RAM. The average processing time for motion estimation per frame was 1.0834 seconds per frame.

Figs. 1(a-l) depict the face model at the position and orientation computed with the estimated motion parameters overlayed with the 13th, 23rd, 30th, 42nd, 60th, 71st, 90th, 100th, 110th, 119th, 128th and 149th frames of the real image sequence Claire. Some tracking errors are visible particularly during the strong rotations of Claire’s head when she looks to the left and down to the table. Small tracking errors are also observed when Claire opens or closes her eyes or mouth. This because the shape and motion models fail on describing the local deformation of these surfaces. Figs. 2(a-g) depict the face model at the position and orientation computed with the estimated motion parameters overlayed with the 1st, 28th, 35th, and 49th frames of the real image sequence Miss America.

Fig. 1. (a-l) Triangular mesh at the position and orientation computed with the estimated motion parameters overlayed with the 13th, 23rd, 30th, 42nd, 60th, 71st, 90th, 100th, 110th, 119th, 128th and 149th frames of the real image sequence Claire.
4. CONCLUSIONS

We have implemented a Maximum-Likelihood motion estimation algorithm and applied it for estimating the motion of a moving human face only. Experimental results showed visible tracking errors particularly when the person looks to the side or down to the bottom. Smaller tracking errors could also be observed when the person opens or closes her eyes or mouth.

5. REFERENCES


Fig. 2. (a-l) Triangular mesh at the position and orientation computed with the estimated motion parameters overlayed with the 1st, 28th, 35th, and 49th frames of the real image sequence Miss America.