Gesture Recognition for Visually Mediated Interaction using Probabilistic Event Trajectories

Stephen J. McKenna † and Shaogang Gong ‡

† Dept. of Applied Computing, University of Dundee,
Dundee DD1 4HN, Scotland. stephen@dcs.qmw.ac.uk

‡ Dept. of Computer Science, Queen Mary and Westfield College,
London E1 4NS, England. sgg@dcs.qmw.ac.uk

Abstract

An approach to gesture recognition is presented in which gestures are modelled probabilistically as sequences of visual events. These events are matched to visual input using probabilistic models estimated from motion feature trajectories. The features used are motion image moments. The method was applied to a set of gestures defined within the context of an application in visually mediated interaction in which they would be used to control an active teleconferencing camera. The approach is computationally efficient allowing real-time performance to be obtained.

1 Introduction

There has been a lot of interest recently within the computer vision community in the recognition of human actions and gestures. This work presents an approach to recognition in which the actions and gestures are modelled probabilistically as a series of visual events. Visual input is analysed by extracting motion feature trajectories. Events are characterised by probabilistic models of feature trajectories estimated from examples. The approach is applied to an application in visually mediated interaction in which pointing and waving gestures are used to communicate commands to an active teleconferencing camera. These gestures can be characterised by coarse spatio-temporal features which are inexpensive to

compute. This property along with their view-specific nature, makes real-time recognition possible.

The feature set used is based on moments estimated from motion images. Image moments have previously been used for gesture recognition. For example, moments extracted from hands tracked using colour were used for American Sign Language recognition [5] and moments based on image axis projections were used in a hand-driven games interface [2]. The method used for matching trajectories is similar in some respects to that used by Black and Jepson to drive a "Condensation" recognition algorithm [1]. The probabilistic method used for treating gestures as sequences of events is similar to a state machine used to parse gestures [6].

2 Gestures for Visually Mediated Interaction

In order to illustrate the approach, let us restrict our attention to a set of four gestures. These gestures have been defined within the context of an application in visually mediated interaction in which they are to be used to control an active camera for teleconferencing. The four gestures are (i) pointing left (ii) pointing right, (iii) waving high up and (iv) waving low down. The associated camera actions for the teleconferencing applications are (i) pan right to next subject, (ii) pan left to next subject, (iii) zoom out to a wide-angle view, and (iv) zoom in to frame the gesturer.

The four gestures are similar in many respects. They are all deictic in the sense that they are each used to communicate a direction of desired motion for the camera field-of-view. The gestures are all motions of the arm and hand in which local spatial information, such as the shape of the hand, is unimportant. The gestures also share a similar temporal structure. They are all 'tri-phasic' i.e. consisting of three phases: (i) an initial transitional phase in which the arm is raised, (ii) a middle phase, and (iii) a final transitional phase in which the arm is lowered to a resting position. The similarity of the gestures has important implications for visual interpretation. While it implies that a unified recognition method should be applicable, it will inevitably mean that ambiguities arise. For example, it is simple to perform a single gesture which human observers will inconsistently classify as both pointing and waving.

An image sequence database was collected containing twelve examples of each gesture. There were three different subjects each of whom performed each gesture four times. Each sequence had 60 frames captured at 12 Hz. An example of each gesture can be seen in Figure 1.

3 Image Motion Moment Features

It is assumed that the subject is sitting relatively motionless and that the major changes in the image sequence are due to the motion of the arm used to perform the gestures. The gestures to be modelled and recognised are by their nature always oriented towards the camera. Therefore, a 2D view-specific representation can be used. Since detailed shape is unimportant, a coarse spatial representation

Figure 1: **Selected frames from four of the sequences used to derive gesture models. The four gestures are (from top to bottom) (i) pointing left, (ii) pointing right, (iii) waving high and (iv) waving low.**

is suitable. In particular, moment features estimated from image motion provide such a representation.

Various levels of sophistication are possible when detecting human motion (e.g. [3]). Here for simplicity and computational efficiency, two-frame temporal differencing was used to detect significant changes in the intensity image. At time t, the binary image $\mathbf{B}_t$ was obtained by thresholding $|\mathbf{I}_t - \mathbf{I}_{t-1}|$ at a value T tuned to the imaging noise (T was set to 10 in our implementation). A set of moment features was then extracted from each image $\mathbf{B}_t$. The motion area A (zeroth-order moment), the centroid co-ordinates $\bar{x}$, $\bar{y}$ (first-order moments) and the elongation E (derived from the second-order moments) are estimated as follows:

$$\begin{aligned} A_t &= \textstyle\sum_{x,y} \mathbf{B}_t[x,y], & \bar{x}_t &= \tfrac{1}{A_t} \textstyle\sum_{x,y} x\mathbf{B}_t[x,y], \\ \bar{y}_t &= \tfrac{1}{A_t} \textstyle\sum_{x,y} y\mathbf{B}_t[x,y], & E_t &= \tfrac{\chi_{max}}{\chi_{min}} \end{aligned} \tag{1}$$

where, $\chi^2 = \frac{1}{2}(a+c) + \frac{1}{2}(a-c)\cos 2\theta + \frac{1}{2}b\sin 2\theta$,

$$a = \sum_{x,y}(x-\bar{x})^2\mathbf{B}_t[x,y], \qquad \sin 2\theta = \pm\frac{b}{\sqrt{b^2+(a-c)^2}},$$

$$b = 2\sum_{x,y}(x-\bar{x})(y-\bar{y})\mathbf{B}_t[x,y],$$

$$c = \sum_{x,y}(y-\bar{y})^2\mathbf{B}_t[x,y], \qquad \cos 2\theta = \pm\frac{a-c}{\sqrt{b^2+(a-c)^2}}$$

The minimum and maximum values of χ are found by changing the signs of the $\sin 2\theta$ and $\cos 2\theta$. Elongation has a lower bound of $E = 1$ in the case of a circularly symmetric motion image. In order to obtain a feature set with invariance to translational shifts in the image-plane, the displacement of the centroid was estimated as $u_t = \bar{x}_t - \bar{x}_{t-1}$, $v_t = \bar{y}_t - \bar{y}_{t-1}$. At time t, the estimated feature set is (A_t, u_t, v_t, E_t).

It is worth pointing out that these features will exhibit variations due to extrinsic factors such as illumination and viewing geometry. In particular, area and centroid displacement will scale differently. The feature set will also vary between different gesturers and between different instances of the same gesture. In particular, the visual texture of clothing will cause variation with more highly textured clothing increasing the motion area.

Figure 2 shows five temporal difference frames from the second sequence in Figure 1. The path described by the estimated centroid during frames with significant motion is also shown. It is clear that during periods of low motion area, estimates of first and second-order moments will become unreliable.

Figure 2: **Selected temporal difference frames from the second sequence in Figure 1. The area of detected motion is shown as mid-grey here, whilst the path described by the estimated centroid is overlaid in black.**

4 Feature Trajectories

A set of feature trajectories can be computed from an observed image sequence by estimating the moment features for each frame. At time t, a temporal trajectory $\mathbf{z}_t = (\ldots, z_{t-2}, z_{t-1}, z_t)$ is available for each feature, where z_t denotes A_t, u_t, v_t or E_t. Different gestures give rise to different feature trajectories and the trajectories for a particular gesture vary between different instances of that gesture.

A probabilistic event model can be obtained from appropriately aligned example sequences of that event. A gesture can be modelled as a series of visual events.

For example, a pointing gesture can be modelled as a sequence of three events corresponding to the phases discussed in section 2.

An event is modelled as follows. Given temporally segmented examples, the mean duration, μ_d, of the event is estimated. Each example's feature trajectories are then temporally scaled to this mean duration using linear interpolation. The aligned examples are then averaged to yield a mean trajectory $(m_1, m_2, \ldots, m_w)$, where w is μ_d rounded to the nearest integer. The variances $(\sigma_1^2, \sigma_2^2, \ldots, \sigma_w^2)$ are also estimated. Therefore, an event model, ξ, consists of a model trajectory $\mathbf{m} = (m_1, \sigma_1^2, m_2, \sigma_2^2, \ldots, m_w, \sigma_w^2)$ for each feature and a duration model $\mathbf{d} = (\mu_d, d_{min}, d_{max})$. The variance parameters play an important role since the variation between examples differs over the length of the event. The probability density for the event duration is modelled as uniform over a finite interval $[d_{min}, d_{max}]$. This was found to be a better model for duration than a Gaussian density $d \sim \mathcal{N}(\mu_d, \sigma_d^2)$.

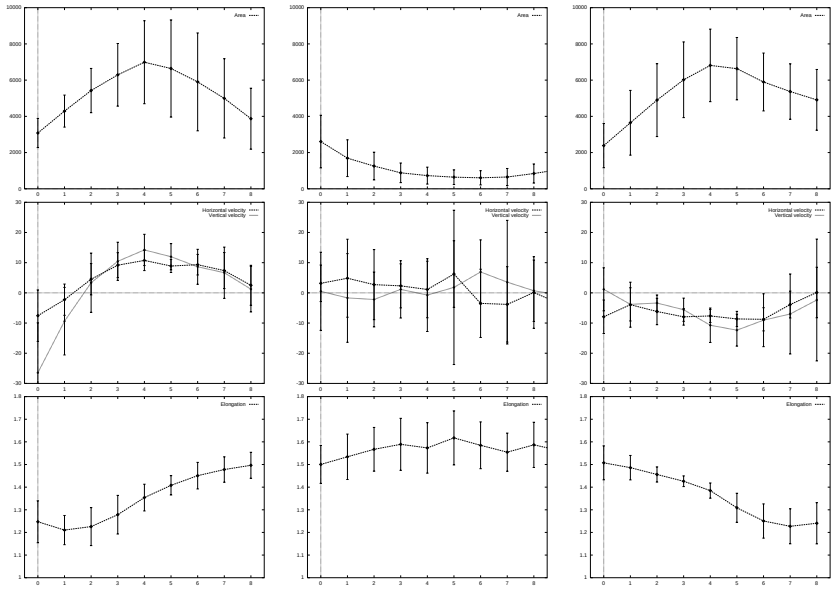

Figure 3: **Model event trajectories for pointing left. From top to bottom: area, displacement and elongation. From left to right: start phase, middle phase and end phase. Error bars denote $\pm\sigma$.**

Figures 3, 4 and 5 show model trajectories estimated from the gesture database described in Section 2. The pointing gestures have been modelled using three events (giving rise to three model trajectories) whilst the waving gestures have been modelled as a single event. This was convenient since the second phase of a pointing gesture contains little significant motion leading to a natural segmentation of the gesture into three visual events. However, segmentation of a waving gesture into three phases is more ambiguous.

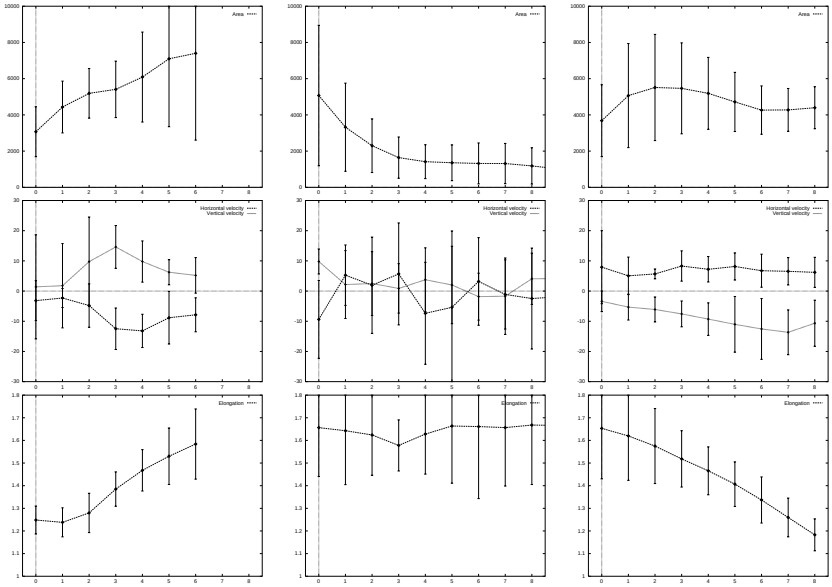

Figure 4: **Model event trajectories for pointing right. From top to bottom: area, displacement and elongation. From left to right: start phase, middle phase and end phase. Error bars denote $\pm\sigma$.**

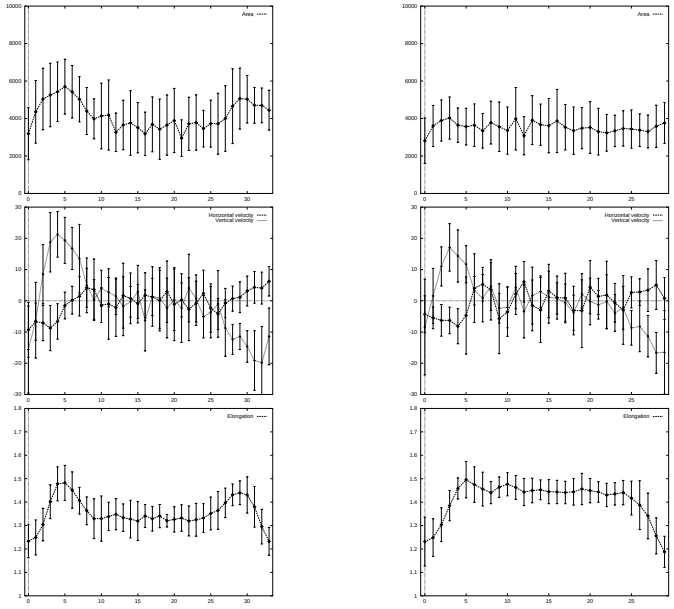

Figure 5: **Model trajectories for high (left) and low (right) waves. Top to bottom: area, displacement and elongation. Error bars denote $\pm\sigma$.**

5 Event Recognition

A w-frame model trajectory $\mathbf{m} = (m_1, \sigma_1^2 \ldots, m_{w-1}, \sigma_{w-1}^2, m_w, \sigma_w^2)$ can be matched with a w-frame trajectory $\mathbf{z}_t = (z_{t-w+1}, \ldots, z_{t-1}, z_t)$ observed at time t to estimate the likelihood $p(\mathbf{z}_t|\mathbf{m})$ of the observed trajectory given the model:

$$p(\mathbf{z}_t|\mathbf{m}) = \frac{1}{\sqrt{2\pi}|\Sigma|^{\frac{1}{w}}} \exp\left(-\frac{1}{2w}\sum_{j=1}^{w}\frac{(z_{t-(w-j)} - m_j)^2}{\sigma_j^2}\right) \tag{2}$$

This likelihood is a Gaussian density function normalised for the length of the temporal window, w. The covariance matrix of this Gaussian is diagonal with elements $(\sigma_1^2, \ldots, \sigma_{w-1}^2, \sigma_w^2)$. The use of diagonal Gaussian matching is motivated by the fact that the variance of features is not consistent within trajectories. This can be seen in Figures 3 and 4. For example, when the motion area is low, the centroid estimates become less stable and have higher variance. It would only be useful to estimate full covariance matrices given a perhaps impractically large amount of training data.
In order to perform matches with time-scaling, a parameter ρ is introduced:

$$p(\mathbf{z}_t|\mathbf{m}, \rho) = \frac{1}{\sqrt{2\pi}|\Sigma|^{\frac{1}{w}}} \exp\left(-\frac{1}{2w}\sum_{j=1}^{w}\frac{(z_{t-\rho(w-j)} - m_j)^2}{\sigma_j^2}\right) \tag{3}$$

Several matches are performed by sampling ρ uniformly in the interval $[\frac{d_{min}}{\mu_d}, \frac{d_{max}}{\mu_d}]$. Given n values of ρ, the likelihood of the observed trajectory given a model is estimated as:

$$p(\mathbf{z}_t|\mathbf{m}, \mathbf{d}) = \frac{1}{n}\sum_{i=1}^{n} p(\mathbf{z}_t|\mathbf{m}, \rho_i) \tag{4}$$

A more robust estimator is:

$$p(\mathbf{z}_t|\mathbf{m}, \mathbf{d}) = \max_i p(\mathbf{z}_t|\mathbf{m}, \rho_i) \tag{5}$$

An event model, ξ, contains a trajectory model for each of the $N = 4$ features, $\xi = (\mathbf{m}_1, \ldots, \mathbf{m}_N, \mathbf{d})$. The likelihood of an event ξ is estimated as:

$$p(\mathbf{z}_t|\xi) = p(\mathbf{z}_t|\mathbf{m}_1, \ldots, \mathbf{m}_N, \mathbf{d}) = \Pi_{i=1}^{N} p(\mathbf{z}_t|\mathbf{m}_i, \mathbf{d}) \tag{6}$$

6 Gesture Recognition

Gestures are modelled as sequences of multiple events. Each event is matched independently with its own event model and linear time-scaling. Recognition of a gesture constitutes matching the appropriate events sequentially. The gesture as a whole is thus time-warped in a piecewise linear fashion.

Gesture recognition is performed using a probabilistic finite state machine. State transitions depend on both the observed model likelihood and the estimated state duration p.d.f. This is similar to a hidden Markov model although we do

not use a transition probability matrix as this essentially models state durations with an exponential favouring shorter durations. Instead the state duration p.d.f. is estimated from the training examples as either a Gaussian $p(d) \sim \mathcal{N}(\mu_d, \sigma_d^2)$ or a uniform density over the interval $[d_{min}, d_{max}]$.

7 Results

Firstly, the 48 sequences in the training database were analysed using a single event model for each of the four gestures. Table 1 shows the results in the form of a confusion matrix. There were 5 errors and these were all due to confusion between high and low waves. All the pointing gestures were correctly identified. While care should be taken when interpreting results on training data, this result would seem to indicate that the models are effective at detecting and discriminating between pointing left, pointing right and waving.

	Point left	Point right	Wave high	Wave low
Point left	12	0	0	0
Point right	0	12	0	0
Wave high	0	0	8	4
Wave low	0	0	1	11

Table 1: Confusion matrix for the training database. Four high waves were misclassified as low waves. One low wave was misclassified as a high wave.

Figure 6 shows the gesture likelihoods obtained by matching the gesture models to a sequence in which a novel subject (not in the training data) points left, waves and points right in that order.

Figure 7 shows the likelihoods of the pointing events for the same sequence.The event models for pointing right also respond to the waving gesture although not strongly enough to result in incorrect recognition. This is intuitively reasonable since these gestures appear similar.

8 Conclusions

Trajectories based on a simple set of image motion features were used to estimate models for gesture events. A temporally normalised Gaussian matching with time-scaling was able to successfully detect and discriminate between pointing left, pointing right and waving gestures. This was despite significant variations between examples of the same gesture, both between gesturers and for a particular gesturer.

Future work will include integration of this approach with methods for active tracking and localisation of heads and hands [4]. The aim is to use the gesture recognition to drive an active camera for teleconferencing.

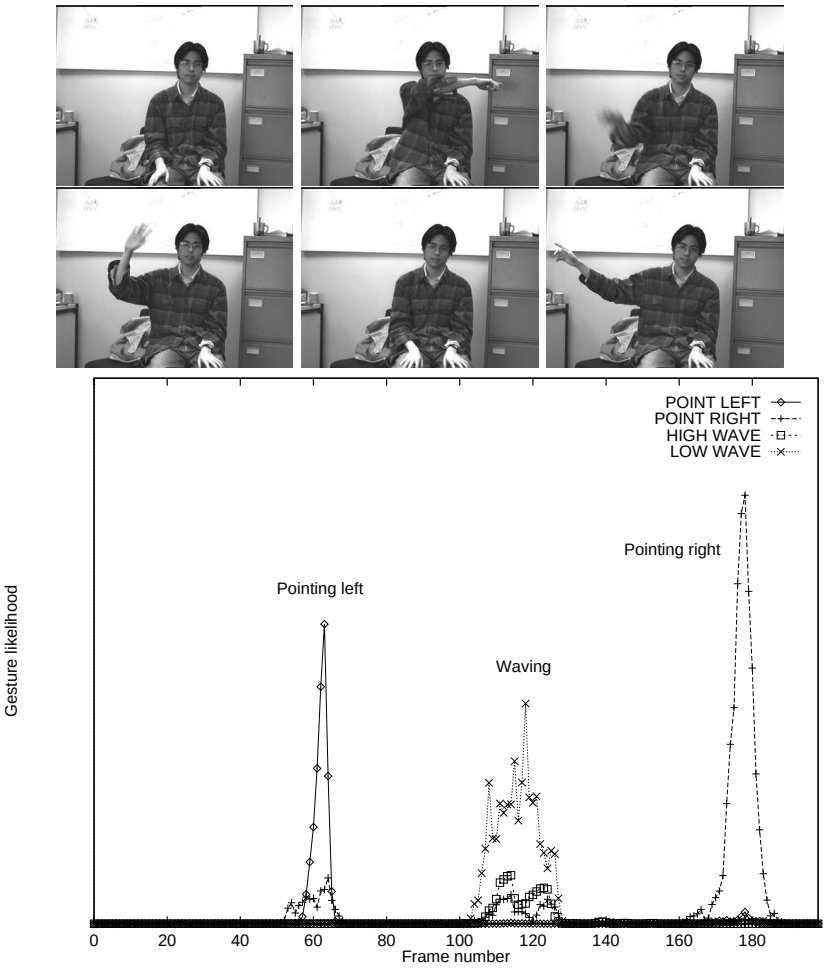

Figure 6: **Top: Frames 20, 50, 85, 100, 135 and 160 from a test sequence in which a novel subject points left, waves and then points right. Bottom: Gesture likelihoods estimated from the test sequence.**

Acknowledgments

S. McKenna was supported by EPSRC grant no. GR/K44657. The authors are grateful to Hilary Buxton and Jonathan Howell for many interesting discussions. They are also grateful to Jonathan Howell for his assistance with the database and to Ong Eng-Jon for volunteering to be a test subject.

References

[1] M. J. Black and A. D. Jepson. Recognizing temporal trajectories using the condensation algorithm. In *Proc. 3nd IEEE Int. Conf. on Automatic Face and*

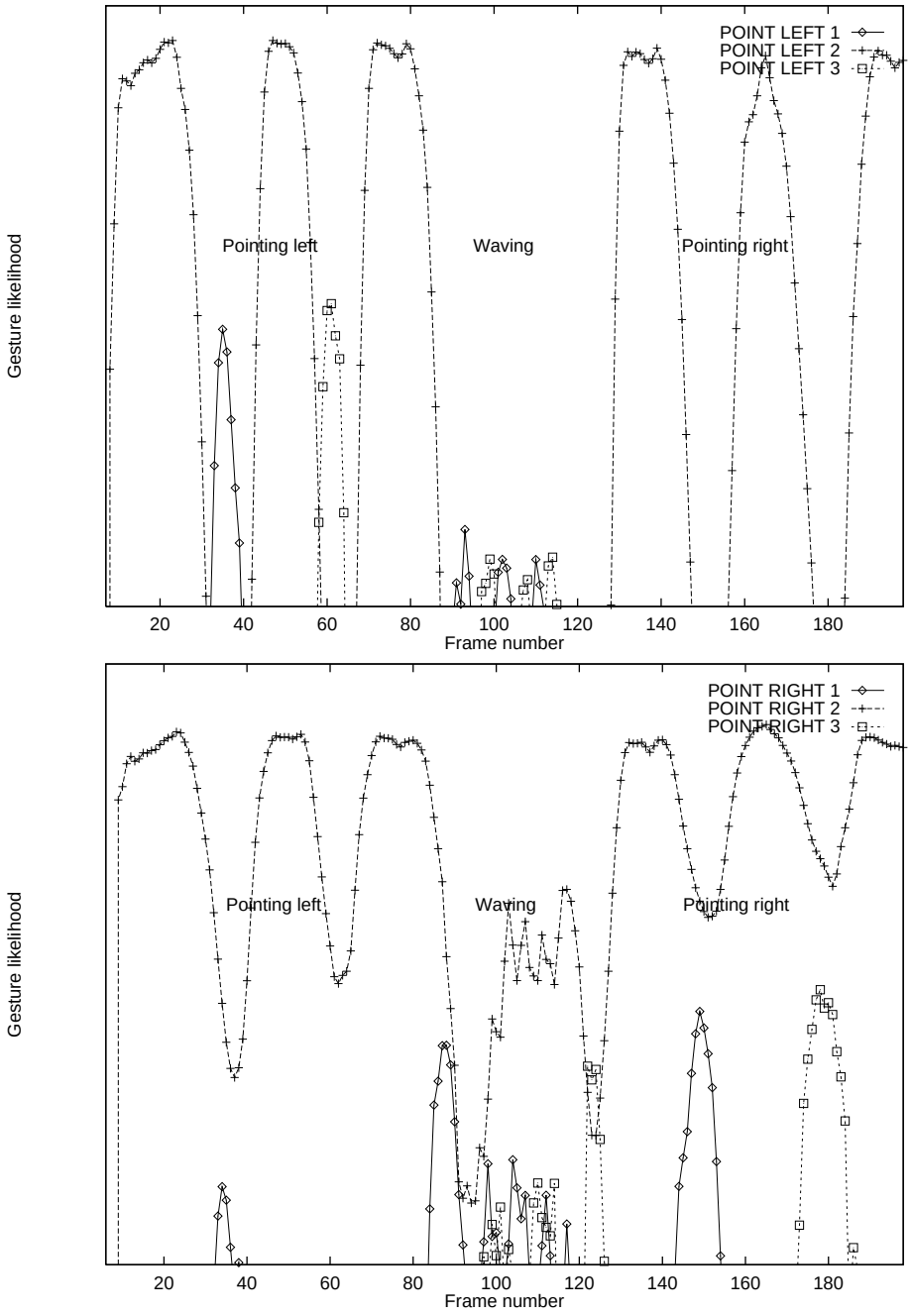

Figure 7: **Event likelihoods estimated from the test sequence in Figure 6. The vertical axis is a logarithmic scale. Top: Likelihoods for the pointing left events. Bottom: Likelihoods for the pointing right events.**

Gesture Recognition, pages 16–21, Nara, Japan, 1998.

[2] W. T. Freeman, K. Tanaka, J. Ohta, and K. Kyuma. Computer vision for computer games. In *Proc. 2nd Int. Conf. on Automatic Face and Gesture Recognition*, pages 100–105, Killington, Vermont, 1996.

[3] S. J. McKenna, S. Gong, and J. J. Collins. Face tracking and pose representation. In *BMVC*, Edinburgh, 1996.

[4] Y. Raja, S. J. McKenna, and S. Gong. Tracking and segmenting people in varying lighting conditions using colour. In *Proc. 3nd IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pages 228–233, Nara, Japan, 1998.

[5] T. E. Starner and A. Pentland. Visual recognition of american sign language using hidden markov models. In *Proc. 1st Int. Workshop on Automatic Face and Gesture Recognition*, Zurich, 1995.

[6] A. D. Wilson, A. F. Bobick, and J. Cassell. Recovering the temporal structure of natural gesture. In *Proc. 2nd Int. Conf. on Automatic Face and Gesture Recognition*, pages 66–71, Killington, Vermont, 1996.