DYNAMIC HAND GESTURE UNDERSTANDING – A NEW APPROACH

M. Yeasin

Electro-Technical Laboratory (ETL)
1-1-4 Umezono, Tsukuba-305, Japan
Email: yeasin@etl.go.jp

S. Chaudhuri

Department of Electrical Engineering
IIT Bombay, Powai, Mumbai-76, India
E-mail: sc@ee.iitb.ernet.in

ABSTRACT

Analysis of a dynamic hand gesture requires processing a spatio-temporal image sequence. The actual length of the sequence varies with each instantiation of the gesture. We propose a novel, vision based system for automatic interpretation of a limited set of dynamic hand gestures. This involves extracting the temporal signature of the hand motion from the performed gesture and is subsequently analyzed by a finite state machine to automatically interpret the performed gesture.

1. INTRODUCTION

The use of hand gestures provides an attractive alternative to cumbersome interface devices for human-computer interaction (HCI). In particular, visual interpretation of hand gestures can help in achieving the ease and naturalness desired for HCI. This has motivated a number of researchers concerned with the computer vision-based analysis and interpretation of hand gestures (for a recent survey of the literature on visual interpretation of hand gestures see [1]). Currently, however, this type of interaction is largely unavailable to intelligent systems. An important new application of machine vision, therefore, is to extend the interface between man and machine, allowing machine to directly perceive what its operator is doing. The ability to follow a hand moving in the space and to recognize a particular motion as a meaningful gesture is, therefore, an essential step in intelligent system design and natural human-machine interaction.

An automatic interpretation of general hand gestures is difficult [2, 3, 4, 5, 6], because it involves analyzing the human hand which has a very high degree of freedom and the mapping of a human gesture onto a particular system function is very difficult. Reasons for this difficulty include individual variations in the exact gestural movement, the problem of knowing when a gesture starts and ends, and variations in the relative positions of other body parts which might help to identify a gesture but are not measured. However, the problem can be simplified with the context of a particular application to develop an appropriate set of gestural commands [7]. In the following section we describe the modeling of dynamic hand gesture to map a subset of gesture to a meaningful system command.

Financial support from COE program funded by STA, ETL, Tsukuba, Japan is gratefully acknowledged.

2. MODELING OF DYNAMIC HAND GESTURE

The gesture is body specific, temporally variable and subject to co-articulation effects [8]. Since human gestures are dynamic processes, it is important to consider the temporal characteristics of gestures. In the constrained case only a small subset of hand gestures is sufficient to interface a machine. Hence, only the following gestures, namely, 'come closer', 'go far', 'move right', 'move left' and 'emergency stop', have been considered for the current implementation. Even in the literature, most of the analyses have been restricted to dealing with a similar set of gestures.

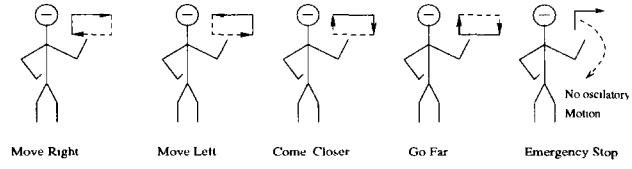

Figure 1: Illustration of a few examples of oscillatory hand gestures. The solid line indicates the start of the gesture along the direction of the arrow and the dashed line indicates the come-back phase of the oscillatory gesture. Looping indicates the oscillatory hand motion. The motion profile of the 'move right' and 'come closer' are opposite to that of 'move left' and 'go far', respectively, and vice-versa. Emergency stop does not have any oscillatory motion.

In order to circumvent various drawbacks of a dynamic hand gesture understanding system, we now examine the basic characteristics of the set of gestures considered here. Let us consider a person performing a 'come closer' gesture by sweeping one hand repeatedly – first quickly towards his body and then slowly away. It may be noted that this typical pattern produces a unique temporal signature. Similarly, in human communication we perform the 'move right' gesture by moving the hand first towards the right direction and then to the left and so forth. We start from an arbitrary spatio-temporal position and perform it. We would like to exploit the temporal signature thus generated for the interpretation purpose. Let us now explain how we define such a temporal signature unambiguously.

Based on the above study we have constructed a gesture lexicon for the subset under consideration and is shown in figure 1. The iconic lexicon shows the motion profile of

the hand (called the knowledge) involved with each gesture and this profile is then exploited for the signature representation and subsequent interpretation purposes. This understanding leads us to construct a set of deterministic finite state machines to represent the temporal signatures, and they satisfy the corresponding motion profiles of the gestures shown in figure 1. By and large these finite state machines are self explanatory and are shown in figure 2. All finite state machines have five states namely, start(S), up(U), down(D), left(L) and right(R). However, only two to three of these states are used to represent a gesture signature. For example, a 'move right' gesture may have a signature of the form 'S-R-L-R-L-R-L'. The self loops are essential to accommodate the idleness of the hand movement while changing the direction of hand waving. This is due to the inertia of motion and would vary from person to person and for each instantiation of the gesture. To further illustrate, the 'come closer' gesture may have the signature 'S-U-D-U-D-U-D'. The length of the signature pattern does not depend on the temporal duration over which the gesture is performed, but depends on the number of changes in the direction of motion for the oscillatory gesture. Given the gesture data the temporal signature extraction would involve processing the image sequence to estimate the dominant direction of motion (such as U, D, L, R) for each subsequence of the temporally segmented data.

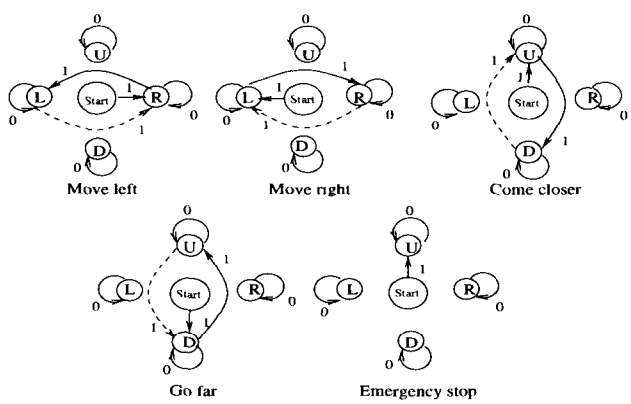

Figure 2: Finite state automaton models constructed for gestures shown in figure 1. A 1 indicates motion and 0 indicates no motion. The solid line indicates the start of the gesture along the direction of the arrow and the dashed line indicates the come-back phase of the oscillatory gesture. Here the states are Start(S), L(left), R(right), U(up) and D(down). Self state transition is used to accommodate the zero motion subsequence that may occur due to the inaptness of the creator during the gesture.

The most desirable part of the finite state modeling is that it makes the system adaptive to the meaning of gesture. Adaptation with the cross cultural gesture can be achieved by redefining the finite state machines according to the thumb rules of the society. The inclusion of new gestures involves simply the construction of additional finite state machines portraying the corresponding motion profile.

3. GESTURE UNDERSTANDING

The essential components of the proposed vision based gesture interpretation system are shown in figure 3. The system has two main modules, namely, i) temporal signature extraction and ii) interpretation of the extracted signature. The first module extracts the temporal signature embedded in the gesture while the interpretation module interprets the extracted temporal signature. In the heart of the temporal signature extraction unit, there should be a motion sensor which can detect the direction and the speed in real time. Unfortunately, currently available methods are not amenable to a real-time implementation. Hence, we seek a faster method to extract the temporal signature (see section 3.1).

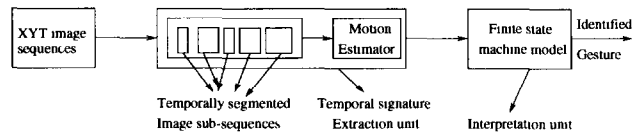

Figure 3: The proposed gesture interpretation system.

The output of a finite state machine shown in figure 2 and described in the previous section responds to the temporal signature of the gesture thus obtained and the identification involves verifying the corresponding production rule of the performed gesture. The introduction of finite state machines makes the system insensitive to the start and the stop positions of the gesture. The key advantage of this approach is that no *training* concept is involved. The system is highly reconfigurable as the inclusion of new gestures needs only to model them suitably using additional finite state machines.

3.1. Temporal Signature Extraction

To extract the temporal signature, the system first temporally segments the gesture data into sub-sequences involving a uniform dynamics, i.e, motion in only one direction (for example, left or right, etc.). Once the temporal segmentation is done we estimate the velocity for each subsequences using an iterative algorithm described in Section 3.1.2. The recovered motion description (U/D/R/L) from all the subsequences are coalesced together to form the temporal signature.

3.1.1. Temporal Segmentation of an Image Sequence

Change detection by subtraction of successive frames is a common practice in motion analysis. Variations of this approach use Laplacian of a Gaussian (LOG) operator before calculating the difference image or use the second order temporal derivative and detect the abrupt change [9]. This can be effectively achieved by convolving the intensity image $I(x, y, t)$ with a LOG operator in the temporal direction yielding the temporal zero crossing $\hat{\mathbf{I}}(x, y, t)$.

$$\hat{\mathbf{I}}(x, y, t) = \frac{\partial^2 \mathcal{G}(t)}{\partial t^2} * I(x, y, t), \qquad (1)$$

where $\mathcal{G}(t) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-t^2}{2\sigma^2}\right)$ and σ is the spread of the Gaussian function. The motion of an edge in the image produces a temporal zero crossing in $\hat{\mathbf{I}}(x, y, t)$ at the location of the edge. Global and temporal intensity changes do not result in such zero crossings. It is sensible to assume that at the motion breakpoint there will be very insignificant motion which implies that ideally there will be no zero crossing at motion break point. Given this background, we proceed as follows:

- Obtain successive difference and temporal zero crossing images for the sequence.
- If there is no zero crossing or the number of zero crossings is less than a threshold, mark the frame for temporal segmentation.

3.1.2. Dominant Motion Estimation Using Motion Energy

Let us consider an object moving with an uniform velocity v_x and v_y in the x and the y directions, respectively, and we are interested in detecting the motion. Such a sequence $I(x, y, t)$ can be described by $I(x, y, t) = I(x, y) * \delta(x - v_x t, y - v_y t)$. In the Fourier domain the relation is given by

$$\hat{I}(f_x, f_y, f_t) = \hat{I}(f_x, f_y)\delta(f_x v_x + f_y v_y + f_t), \quad (2)$$

where f_x, f_y and f_t are the spatial and the temporal Fourier variables, respectively and $\hat{I}(f_x, f_y)$ is the Fourier transform of $I(x, y, 0)$. Equation (2) implies that an object moving with an uniform velocity occupies only an $n-1$ dimensional space in the n dimensional space in Fourier domain. In a two or three dimensional space, it is a line or a plane, respectively. The equation of the plane is directly given by the argument of the δ function

$$f_t = -(f_x v_x + f_y v_y). \quad (3)$$

Based on the above observation an interesting set of filters originate from the models used to describe motion in biological visual systems. Gabor-like quadrature filters are used to determine the image motion from which the term *motion energy* [10] is coined. We would like to emphasize that in our problem (temporal signature extraction) the idea is to only detect the direction of dominant motion (spatio-temporal orientation). The basic idea is to formulate the problem in an optimization framework and we use a gradient descent method to arrive at the desired solution. We use the motion energy itself as the cost function to get an estimate of center frequencies $f_{x_0}, f_{y_0}, f_{t_0}$ where the motion energy is maximum. Hence, the cost function can be written as

$$C(f_{x_0}, f_{y_0}, f_{t_0}) = \sum_{f_x, f_y, f_t} \left|\hat{I}(f_x, f_y, f_t)\right|^2 G(f_x, f_y, f_t; f_{x_0}, f_{y_0}, f_{t_0}). \quad (4)$$

The center frequency of the band pass filter for which the motion energy is maximum can be found iteratively. Once the filter center frequencies are known the dominant motion (local orientation) of the signal in frequency domain is known. We use the above algorithm to estimate dominant motion of each sub-sequence to obtain the temporal signature which was subsequently used by the deterministic finite state machine to interpret the gesture.

4. EXPERIMENTAL RESULTS

The computational representation scheme proposed in this paper involves temporal signature extraction from a performed gesture. A few sample frames from 'move right' and 'come closer' gestures are shown in figure 4 for illustration.

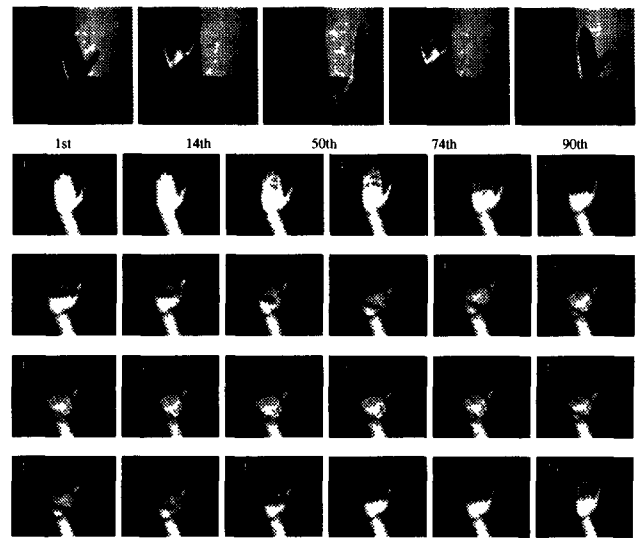

Figure 4: The top row shows a few sample frames from the 'move right' gesture. The 'move left' gesture is opposite to 'move right' gesture and is not shown. Below, one cycle of the 'come closer' gesture is shown. The 'go far' gesture is similar.

The first step in temporal signature extraction involves finding the motion breakpoint of the gesture. The application of the proposed temporal segmentation algorithm (see section 3.1.1) detects the motion break point. Figure 5 shows the result of detecting motion breakpoint for 'come closer' gesture (see figur 4 for the corresponding data). It depicts the detected temporal zero crossings in all frames (only one cycle has been displayed for the sake of clarity). Larger the motion, higher is the number of zero crossings. Figure 6 shows the plot of number of frame - number of zero-crossings. From figures 5 and 6 it is evident that frame number 15 has a very few zero crossings and all the other frames has quite a good number of zero crossings. After thresholding we declare this frame as the motion break point. This implies that frames $0-14$ corresponds to a temporally segmented sub-sequence displaying a motion sequence along a particular direction (down or closing the hand). Similarly, frames $16-24$ corresponds to another subsequence displaying up motion or opening the hand.

The motion estimation scheme described in the previous section is used to estimate the average velocity and direction of such temporally segmented motion sub-sequences. Experimentally we found that the proposed scheme yields correct estimates when they are provided with good initial conditions. This is due to the fact that in a real image

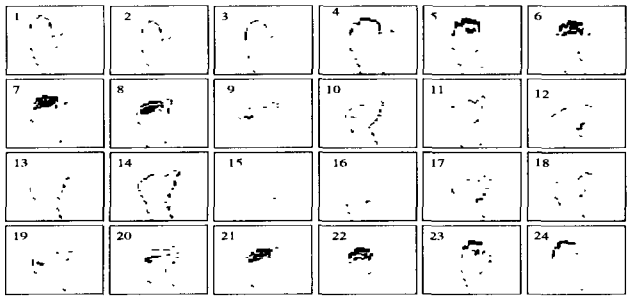

Figure 5: Temporal zero crossings images obtained for the 'come closer' gesture. Only 24 frames from a image sequence of 96 frames has been shown for the sake of clarity. Frame number 15 has been marked for motion breakpoint.

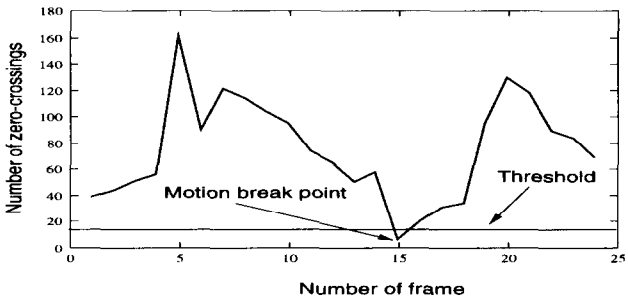

Figure 6: Plot of number of zero-crossings against the number of frame for the 'come closer' gesture. It is clearly evident that the frame 15 corresponds to the motion breakpoint.

sequence the motion is hardly uniform. The motion parameter (direction information) thus obtained from each subsequence is coalesced together to form the temporal signature. The extracted temporal signature obtained for the above 'come closer' gesture is S-U-D-U-D-D-U-D-U-D-U-U-D and for the 'move right gesture' is S-R-L-L-R-R-L-R-L-R-L-L-R-R-L, where 'S' indicates start of the gesture, 'U', 'D', 'R', 'L' indicates up, down, right and left-ward motion, respectively. We obtain similar results for all the other gesture sequences. In order to provide a good initial estimate of the center frequencies of the bandpass filters, we run the optimization algorithm several times with different initial guesses and select the direction which results most frequently. Having obtained the temporal signature for a test gesture, we verify which production rule for a particular gesture is satisfied. The test gesture is now identified by matching the production rule. In all our experiments we identified all test gestures correctly.

5. CONCLUSIONS

Understanding of dynamic hand gestures requires an analysis of spatio-temporal image sequences. The actual length of the sequence varies with each instantiation of the gesture. The key motivation to solving the problem is to translate the richness of human gestural communication power to a computer for a better HCI. To find the velocity and hence to extract the temporal signature, our effort involved an optimization of the motion energy. The proposed iterative scheme is computationally efficient compared to the traditionally used quadrature filter based methods, albeit far from being real time. The proposed motion estimation scheme is sensitive to the choice of the initial guess for real image sequences. The finite state machine modeling of the dynamic hand gesture helps in interpreting the gesture accurately and also avoids the computationally intensive task of image sequence warping. As we work with monocular image sequences and the optimization scheme is not prohibitively demanding, the approach may be suitable even for a moderate hardware setup. Additionally, our formulation does not impose any condition on start and stop positions of the gesture. The system is highly reconfigurable as the inclusion of a new gesture needs only to model it appropriately by an additional finite state machine.

6. REFERENCES

[1] V.I. Pavlovic, R. Sharma and T.S. Huang, "Visual interpretation of hand gestures for human-computer interaction: A review", *IEEE Tran. on Pattern Analy and Mach. Intell.*, vol. 19(7), July, 1997.

[2] M. Yeasin and S. Chaudhuri, "Automatic generation of robot program code : Learning from perceptual data", *To appear in the proceedings of ICCV'98, Bombay, India*, 1998.

[3] T. Ahmad, C.J. Taylor, A. Lanitis and T.F. Cootes, "Tracking and recognising hand gestures, using statistical shape models ", *Image And Vision Computing*, vol. 15(5), pp. 345–352, May, 1997.

[4] D.J. Sturman and D. Zeltzer, "A survey of glove based input", *IEEE Computer Graphics and Application*, vol. 14, pp. 30–39, 1994.

[5] S.S. Fels and G. Henton, "Glove-Talk: A neural network interface between Data-glove and speech synthesizer", *IEEE Trans. on Neural Network*, vol. 4, pp. 2–8, 1993.

[6] R. Koch, "Dynamic 3D scene analysis through synthetic feedback control", *IEEE Tran. on Pattern Analy and Mach. Intell.*, vol. 15(6), pp. 556–568, 1993.

[7] V.I. Polvic, R. Sharma and T.S. Huang, "Gestural computing interface to visual computing environment for molecular biologists", in *Intl. Conf. on. Automatic Face and Gesture Recognition*, Vermont. USA, Oct. 14-16, 1996.

[8] M. Brand and T. Darrell, "Cuasal analysis for visual gestures understanding", Tech. Rep., MIT media lab. Tech. Report, TR-327, 1995.

[9] W. Chen and N. Nandhakumar, "A simple scheme for motion boundary detection", *Pattern Recognition*, vol. 29(10), pp. 1689–1701, 1996.

[10] D.J. Hegger, "A method for extraction of image flow", *Journal of Optical Society America*, vol. A2-4(8), pp. 1456–1471, 1987.