Vision Based Hand Gesture Interpretation Using Recursive Estimation

Jennifer Schlenzig* Edward Hunter Ramesh Jain

Visual Computing Laboratory
University of California, San Diego
La Jolla, CA 92093-0407

Abstract

Gesture recognition requires spatio-temporal image sequence analysis. The actual length of the sequence varies with each instantiation of the gesture, and can be quite long in the case of a multiple gesture sequence. To achieve adequate system response we introduce the concept of recursive estimation of the gesture state. This consists of modeling the gestures as a sequence of static hand poses. Using a hidden Markov model where the unobservable state is the spatio-temporal gesture and the hand poses are the observations allows us to determine the current probablilities of each gesture with a finite state estimator. This decomposes the gesture recognition process into two stages: identification of the hand pose within the current image frame and incorporation of the new information into the probability estimates. We illustrate the performance of the estimator by describing the implementation of a telerobotic application.

1 Introduction

Providing a computer with the ability to interpret human hand gestures is a step toward more natural human-machine interactions. Existing input systems augmented with this, as well as such other human-like modalities such as speech recognition and facial expression understanding, will add a powerful new dimension to the range of future computer applications and the accessibility of existing ones. A wide spectrum of research is underway on the problem of gesture interpretation (e.g., [1, 2, 3]), but currently there is no universal definition of what a gesture recognition system should do or even what is a gesture.

Our definition of a gesture from the perspective of the computer is simply a temporal sequence of images of the hand. An element from a finite set of static hand poses is the expected content within an image frame. A gesture is, therefore, a sequence of static hand poses. Poses are assumed to contain the identity of the hand shape and (possibly) the orientation, translation and distance from camera information. The spatio-temporal nature of the gesture data make the gesture state unmeasureable at a given instance in time, but for each timestep we can determine the static hand pose. The gesture data and the pose data are two stochastic processes which possess the necessary requirements to allow us to use a hidden Markov model (HMM) [4] to describe the system. Specifically, the gesture data is an unobservable random sequence whose behavior can be summarized with a state transition matrix consisting of the probabilities of each state occurring given only the previous state. The pose information is the observation sequence whose output depends on the current gesture. Because of the dependence of the pose sequence on the gesture data, we can infer what gesture is occurring based on the observed sequence of poses.

Previous research [2, 5, 6] on the use of HMMs to model motion information demonstrates the highly accurate results which may be achieved. These implementations require that each possible motion be modeled with a HMM. Classification of the motion then consists of determining the HMM most likely to have generated the observation sequence. The resulting performance of the classifier is outstanding, but several limitations exist. First, at the time of classification, the entire sequence must be available. This introduces significant I/O latency into the gesture recognition system. Also, the classification of sequences containing more than one motion or gesture is difficult. It requires either the decomposition of the sequence into subsequences or the design of HHMs to represent all possible combinations of multiple gestures. The application of a finite state estimator to the problem of identifying the HMM modeled gesture overcomes both of these limitations.

We impose several constraints on the behavior of the gesture recognition system. First, the system must be unemcumbering. Gloves and tethers interfere with the use of the hand for other tasks. Gloves also require a conscience preparation (putting on the glove) and

*Corresponding author, e-mail address: jschlenz@ucsd.edu

1058-6393/95 $4.00 © 1995 IEEE

further, may impose the need for a personal glove to be available for each user. In addition, we require that the system provide the user with continuous feedback indicating what gesture is currently being performed. This prohibits the use of batch operations which require an entire image time sequence to be available before the identification of the gesture can occur. Benefits of feedback include the fact that subliminal training can take place as an operator naturally alters his motions to compensate for misidentifications. The information provided to the operator can also be exploited by the gesture recognition system, allowing it to modify the type of processing that is performed based on its interpretation of the gesture state. Finally, the more obvious characteristics which we require of the gesture recognition system is that computational requirements must permit real-time interpretation and a significant number of gestures must be robustly indentified.

This paper describes the application and implementation of recursive finite state estimation to the problem of gesture interpretation. We first present a statement of the gesture recognition problem in terms of recursive estimation. This is followed by the discussion of the implementation of a gesture recognition system for the remote control of a robot. Lastly, we summarize the issues involved in recursive estimation.

2 Gesture estimation

The analysis of image sequences can be greatly hindered by the large amount of data being processed and stored. Well-behaving single image algorithms such as template matching are difficult to scale up. As the sequences become longer batch operations will perform with intolerable latency. The alternative to batch operations is the use of recursion. Recursion allows us to incorporate new measurement information into our analysis without having to reprocess previously obtained data, and can be expressed mathematically as

$$\hat{x}_n = \hat{x}_{n-1} + \gamma_n \nu_n$$

where $\hat{x}_n$ is the estimate of the state of the system at time n, ν_n is the new information, or *innovations*, and γ_n is the weighting factor.

At each time step, the innovations is obtained from the current image frame, and is used to modify the estimate of what gesture is occurring. The estimate is then given to the operator informing him of the system's interpretation. The system can use the current estimate to initiate its reaction before the completion of a gesture. This is significantly different from a batch operation which outputs information only at the end of each sequence.

By using an HMM we decompose the problem into two disjoint stages. The first stage is the extraction of the pose information from each image frame as it is captured. This is followed by the incorporation of the new measurement into the gesture estimate. In other words, the image processing portion of the sequence interpretation system has been reduce from temporal analysis to single frame (or when motion segmentation is used, a few frames) processing. This allows us to use well-established techniques in the determination of the hand pose (see for example [7]). The temporal information is not neglected. Instead, the historical data is summarized by the previous estimate value.

2.1 Pose identification

Identifiying the hand pose requires segmenting the object from the background, normalizing it with respect to translation and scale, extracting discriminatory features and applying a decision rule to the feature set for classification. Segmentation techiques which exploit motion information [8, 9] are appropriate because we can assume the hand is the fastest moving object in the image.

Looking beyond the need for insensitivity to noise and preprocessing errors, our choice of a feature set is influenced by two key feature criteria we believe are necessary for robust pose classification: **rotational invariance** and **orthogonality**.

The problem of robust pose classification becomes more tractable if the task of hand shape identification is separated from the determination of hand orientation. This is accomplished by exploiting features which are invariant to image rotation about the centroid. The result is a reduced number of distinct shape classes, which both simplifies the classifier design and improves its performance. The separation of these tasks is further justified in gesture applications where the orientation information is of a coarse nature. Pointing at a quadrant of the computer screen, directing a robot to go left, and interpretation of sign languages are all examples where large rotational variations amongst different instantiations of particular poses are common. In addition, some poses contain no orientation information, and must be similarly classified regardless of their rotation. In these cases, having distinct classes for each orientation needlessly obscures the feature space. Furthermore, pose s which do include orientation should interpret it in a manner dictated by the pose class. For example, a "direct manipulative" gesture (e.g., pointing to a requested destination) requires a more accurate spacial interpretation than a symbolic gesture (e.g., waving hello).

The use of orthogonal features allows us to easily identify optimal subsets of features, which provide the majority of the discriminatory information. It is likely that we would have to use a much larger set of

nonorthogonal features to get the same discriminatory information as in a small set of orthogonal features. In effect, orthogonality allows us to throw away features that contain little discriminatory information. In considering which features to retain, we use the measures of interclass and intraclass distances. Under ideal conditions, features of a given class will always occur (i.e., with zero variance) at a particular value. This value can be taken as the sample mean of the features determined experimentally. Discriminatory ability can be expressed as proportional to the distance between these sample means (interclass distance), and inversely proportional to the variance in the features (intraclass distance)

$$J_i = \frac{Var_j(E_i\Omega_{ij}))}{E_j(Var_i\Omega_{ij}))} \quad (1)$$

where Ω_{ij} represents the i^{th} example and the j subscript denotes that the outer statistics are taken over the j class means and variances. This quantity is computed experimentally (over a training set) for all features, which are then ranked in descending value. The highest ranking features yield the basis for classification. For the case $p > 2$, this measure does not guarantee an optimal subset of features for discriminating all of the classes because particular pairs of classes may overwhelm the feature subset with many highly discriminating features, thus preventing other pairs from getting important features near the top of the list. For the case $p = 2$, this measure can be considered optimal. Therefore, equation 1 is computed for all distinct pairs of classes, and the best features from each pairwise evaluation is included in the subset used for classification.

2.2 The estimator

Once the poses and gestures have been defined for a given application the parameters of the HMM can be determined. In particular, each gesture i yields a preliminary state transition matrix q^k which contains the pose transition probabilities such that $q^k_{ij} = P\{p_j[n]|p_i[n-1]\}$ is the probability of changing from pose i to pose j and the diagonal elements are determined by the normalization constraint: all rows must sum to zero. Note that these probabilities can easily be determined from experimental data. An addtional preliminary transition matrix, q, contains the gesture transition probabilities. Typically, all gesture are assumed to be equally probable. The preliminary transition matrices are combined using,

$$Q = (q \otimes I_L) + diag(q^k; k \in K) \quad (2)$$

where L is the number of poses, K is the number of gestures, I_L is the $L \times L$ identity matrix, $\otimes$ is the kronecker product and diag implies block diagonal, to form a single transition matrix [10].

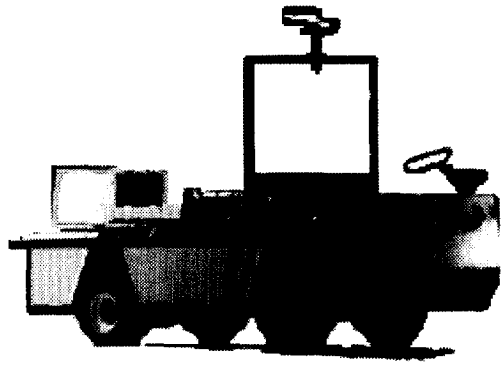

Figure 1: Viscomp Lab's outdoor autonomous vehicle

The estimator, $\hat{\phi}_n$ is assumed to have the form

$$\hat{\phi}_n = E\{\phi_n|Z_n\} \quad (3)$$

where ϕ_n is an indicator vector describing the true gesture information and Z_n is the measurement data available at time n. The derivation of the estimator [11, 12, 13] yields the update equation

$$\begin{aligned} d\hat{\phi}_n &= Q'\hat{\phi}_n\,dt + P_{\phi\phi}\lambda D' R_n^{-1}\,d\nu_n \quad (4)\\ P_{\phi\phi} &= \phi_n\hat{\phi}'_n - \hat{\phi}_n\hat{\phi}'_n \\ R_n &= diag(\lambda D\hat{\phi}'_n) \end{aligned}$$

where the first righthand term in equation 4 is the affect of the model and the second term provides the change due to the current measurement. In equation 4, $d\nu$, is the vector e_i where i is the current pose symbol, λ is the image frame capture rate and $dt = 1$. The discernibility matrix, D, contains the probabilities describing the observation process.

3 An example

To demonstrate the performance of the recursive gesture recognition system we have implemented a tele-robotic application. Working in conjunction with the UCSD Autonomous Outdoor Robotics group we designed a gesture interpretation which allows us to remotely control a robotic gopher (figure 1). The vehicle operates at constant speed while the operator is responsible for steering.

For this application, all of the gestures were time sequences of a single pose. The use of the estimator is still beneficial due to its ability to absorb time variations in the instantiations of the gesture and its graceful handling of pose misclassifications. The pose/gesture set {fist, open 5, point, thumb, ell, hole} is illustrated in figure 2). The gestures were performed

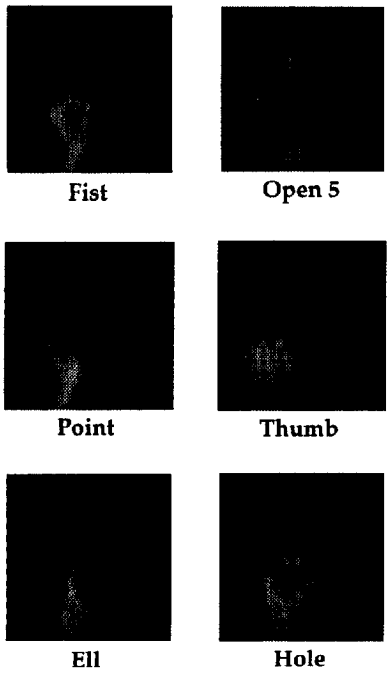

Figure 2: Poses used in this example

against a controlled background enabling us to segment the hand shape via adaptive thresholding.

Our features were the Zernike moments of the binary silhouette images of the hand, cropped at the wrist. Teague [14] introduced an orthogonal, rotationally invariant set of two-dimensional moments based on the Zernike polynomials [15]. They are written as

$$A_{nl} = \frac{n+1}{\pi} \int_0^{2\pi} \int_0^1 (V_{nl}(r,\theta))^* \cdot f(r\cos\theta, r\sin\theta)\, r\, dr\, d\theta \tag{5}$$

valid for $n - |l| = \text{even}$, $|l| \leq n$, with Zernike polynomials defined as

$$V_{nl}(x,y) = V_{nl}(r\cos\theta, r\sin\theta) = R_{nl}(r)\exp(il\theta) \tag{6}$$

$$R_{nl}(r) = \sum_{s=0}^{(n-|l|)/2} \frac{(n-s)!}{s!\left(\frac{n+|l|}{2}-s\right)!\left(\frac{n+|l|}{2}-s\right)!} \tag{7}$$

The magnitudes of these moments, $|A_{nl}|$, which have been successfully used for character recognition [16], have the characteristics we require in a feature set as justified by the following facts.

The Zernike moments of an image and those of the same image rotated through angle ψ are related only by a phase factor $exp(-il\psi)$. Thus the magnitudes of the Zernike moments have the desired property of rotational invariance.

Zernike polynomials possess the property that they are orthogonal to one another over the unit disk. Thus

Figure 3: An example (containing preprocessing errors) of a binary image of the hand and the image reconstructed using the extracted Zernike moments through $order = 25$.

the moments defined above can be individually graded according to equation 1, and an optimal feature subset is easily found. Previous work has compared orthogonal and nonorthogonal moments in detail [17, 18, 19].

Reconstruction of image data from its features is widely regarded as a good measure of the representation ability of the feature set. Intuitively, a feature set that can provide an accurate reconstruction of the original data is seen to provide more discriminatory information in classification than one that can not (see figure 3).

For this application, our classifier was a feedforward neural network trained using conjugate gradient search. The topology of the net consisted of 10 input units, two hidden layers containing 9 and 7 units, and six output units. We trained the net using 100 examples of each pose for only 1007 epochs. This yielded a classifier with an accuracy of 96% on a test set containing as additional 100 examples of each pose.

The parameters of the estimator were determined by experimentation. The diagonal elements of the state transition matrix, Q, were found by taking the inverse of the average execution time of the gesture and mulitplying it by -1. In this case, all gestures were expected to last 2 time steps. The discernibility matrix, D, is telling the estimator that pose classification is only correct 50% of the time, and the error is evenly spread other all other poses. This is exceedingly pesimistic based on the neural network testing, but was chosen due to the fact that the actual usage of the system takes place in an environment much different from the laboratory setting present during the collection of the training data.

In the end the system operated at a rate of $\frac{1}{2}Hz$ with an accuracy which allowed the operator to easily manuever the vehicle within a confined environment. This included snaking through cement posts spaced fifteen feet apart.

4 Conclusion

Gesture interpretation has become recognized as a necessary component of an intuitive human-computer interface. Potential applications include tele-robotics, scientific visualization and sign language interpretation. The use of recursive estimation yields a system which can provide the user and system with gesure information at each timestep. This information can be exploited in several ways including non-intrusive training of the operator. It can also be used to reduce the computational and digitial storage requirements of the system.

References

[1] J. Davis and M. Shah, "Gesture recognition," Tech. Rep. CS-TR-93-11, University of Central Florida, 1993.

[2] J. Yang, Y. Xu, and C. S. Chen, "Gesture interface: Modeling and learning," in *Proceeding of the IEEE Conference on Robotics and Automation*, April 1994.

[3] T. J. Darrell and A. P. Pentland, "Space-time gestures," in *Proceeding of the Conference on Computer Vision and Pattern Recognition*, June 1993.

[4] L. R. Rabiner and B. H. Huang, "An introduction to hidden markov models," *IEEE ASSP Magazine*, pp. 4–16, January 1986.

[5] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden markov models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 379–385, June 1992.

[6] S. R. Veltman and R. Prasad, "Hidden markov models applied to on-line handwritten isolated character recognition," *IEEE Transactions on Image Processing*, pp. 314–318, May 1994.

[7] A. K. Jain, *Fundamentals of Digital Image Processing.* Englewood Cliffs, NJ: Prentice Hall, 1989.

[8] S. H. Haynes and R. Jain, "Detection of moving edges," *Computer Vision, Graphics, and Image Processing*, pp. 345–367, March 1983.

[9] R. Jain and H.-H. Nagel, "On the analysis of accumulative difference pictures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 206–214, April 1979.

[10] D. D. Sworder, R. Vojak, and R. G. Hutchins, "Gain adaptive tracking," *Journal of Guidance, Control, and Dynamics*, pp. 865–873, September-October 1993.

[11] V. Krishnan, *Nonlinear Filtering and Smoothing: An Introduction to Martingales.* New York, NY: John Wiley and Sons, 1984.

[12] D. D. Sworder, "Tactical decision making under stress," Tech. Rep. 0083, Naval Ocean Systems Center, 1991.

[13] J. Schlenzig, E. Hunter, and R. Jain, "Recursive identification of gesture inputs using hidden markov models," in *Proceeding of the Second Annual Conference on Applications of Compter Vision*, December 1994.

[14] M. R. Teague, "Image analysis via the general theory of moments," *Journal of the Optical Society of America*, pp. 920–930, August 1980.

[15] A. B. Bhatia and E. Wolf, "On the circle polynomials of zernike and related orthogonal sets," *Proceedings of the Cambridge Philosophical Society*, vol. 50, 1954.

[16] A. Khotanzad and Y. H. Hong, "Rotation invariant image recognition using features selected via a systematic method," *Pattern Recognition*, vol. 23, pp. 1089–1101, 1990.

[17] C.-H. Teh and R. T. Chin, "On image analysis by the method of moments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 496–513, July 1988.

[18] Y. S. Abu-Mostafa and D. Psaltis, "Recognitive aspects of moment invariants," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 698–706, November 1984.

[19] R. J. Prokop and A. P. Reeves, "A survey of moment-based techniques of unoccluded object representation and recognition," *CVGIP: Graphical Models and Image Processing*, pp. 438–460, September 1992.

[20] M. K. Hu, "Visual pattern recognition by moment invariants," *IRE Transactions on Information Theory*, pp. 179–187, January 1962.

[21] M. Umeda, "Recognition of multi-font printed chinese characters," in *IEEE Conference on Pattern Recognition*, pp. 793–796, 1982.