GESTURE RECOGNITION USING POSITION AND APPEARANCE FEATURES

Tushar Agrawal

Analog Devices Inc.
Mumbai, India.
tushar.agrawal@analog.com

Subhasis Chaudhuri

IIT Bombay,
Mumbai, India.
sc@ee.iitb.ac.in

ABSTRACT

In this paper a scheme for recognizing hand gestures is presented using the output of a Condensation tracker. The tracker is used to obtain a set of features. These features consisting of temporal evolution of the spatial moments form high dimensional feature vectors. The principal components of the feature trajectories are used to recognize the gestures.

1. INTRODUCTION

Hand gesture recognition is essential in a host of applications like haptic interfaces for large-screen multimedia and virtual reality environments [1], robot programming by demonstration, sign language recognition, human computer interaction, telerobotic applications,etc. Previous attempts at recognizing similar gestures have used a variety of methods. Davis and Shah [2] used markers for tracking finger tips and used the fingertip trajectories for recognizing seven gestures. In [3], the hand's position in the image, velocity and values obtained by eigen analysis are used as features to recognize the sign language (SL) using an HMM based scheme. In [4], 53 signs from SL are recognized using 3D data of the arm shape obtained from 3 cameras. In [5], the concept of motion energy is used to estimate the dominant motion of the hand and the gestures are recognized by fitting finite state models of gestures. In [6], the Mahalanobis distance between feature vector consisting of various Hu moments [7] of motion energy image and motion history image of the input and the moment description of each of the known actions are compared. In [8], affine transformations of the hand region are obtained from one frame to the next. This is used to calculate pixel trajectories which are fed to a time delay neural network for recognition. In [9] gestures are modeled as sequences of visual events which are matched to the probabilistic models estimated from feature trajectories. In [10], a gesture is classified as a sequence of postures using principal component analysis and recognized using finite state machines. Multiple cameras are used in [11] to extract the 3D pose of the human body. Instead of using all the parameters describing the pose as features, the trajectory obtained by a projection into a 2D eigenspace is used for gesture recognition. In [12], each feature trajectory is split into sub-trajectories and the recognition is achieved by maximizing the probability of it being a particular gesture in the eigenspaces of each these sub-trajectories. The first few eigenvectors computed from an image sequence are used as features in [13] and each gesture is represented as a sequence of fuzzy states which are recognized using a dynamic programming algorithm. An incremental recognition strategy that is an extension of the condensation algorithm is proposed in [14] to recognize gestures based on the 2D hand trajectory. Gestures are modeled as velocity trajectories and the condensation algorithm is used to incrementally match the gesture models to the input data. In [15] 2-D motion histogram is used as gesture signature and recognition is achieved by comparing a suitable distance metric between gesture signatures.

The developed system tracks a rectangular window bounding the hand in order to extract features which can be used for recognizing the gestures. We detect skin colored blobs in the tracked window and evaluate various order spatial moments of these blobs. These features are calculated for each frame of gesture which results in a feature (temporal) trajectory. Thus we have a trajectory corresponding to each feature. To avoid the problem of each gesture being of different length in time, we normalize the gesture trajectories in time between [0:1]. We then use principal component analysis [16] to classify the feature trajectories into known gestures classes.

2. SYSTEM OVERVIEW

The proposed gesture recognition system is intended to be used as an interface for a telerobotic system. The block diagram of the proposed system is shown in Figure 1.

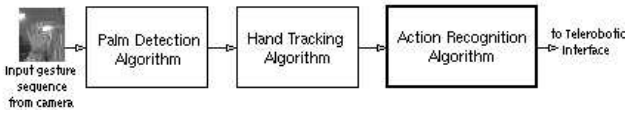

Fig. 1. Block diagram: Gesture Recognition system.

This work is part of Tushar's masters thesis at IIT-Bombay.

User performs a gesture which is captured by the camera. This video sequence is fed as an input to the palm detection algorithm of [17]. The palm in the first frame is detected based on matching a line representation of edges detected in an image with idealized line templates for fingers. The position of the palm in the first frame is fed as an input to the hand tracking algorithm of [18]. Here we use conditional density propagation over time. It is a factored sampling approach to propagate the entire probability distribution of the parameters to be tracked over time. We track the parameters of the rectangular box bounding the hand. The output of the hand tracker is given as input to the recognition algorithm which is the central topic of this paper. The gesture is classified as one of the eight gestures in the vocabulary as would be described in the later sections. The recognized gesture is fed to the telerobotic interface.

3. FEATURE SELECTION

We plan to operate a robot by performing hand gestures which will be recognized as a command by the robot. Dynamic manipulative hand gestures are best suited for such applications. We select the eight gestures namely *'Move Right'* (RIGHT), *'Move Left'* (LEFT), *'Move Up'* (UP), *'Move Down'* (DOWN), *'Move Counterclockwise'* (CCW), *'Move Clockwise'* (CW), *'Move Away'* (AWAY) and *'Move Closer'* (CLOSER) to form our gesture vocabulary. These gestures are characterized both by the motion of the hand in space and the variation in the shape of the hand as the gesture is performed. Examples of few gestures can be seen in Figure 2. Hence in order to recognize gestures, we need to have features which can capture these gesture characteristics. The hand tracker mentioned in the previous section tracks the change in positionand shape of a rectangular window bounding the palm region of the hand performing the gesture. Hence various order spatial moments of the hand region during the motion should serve as good features as temporal evolution of these spatial moments captures the variation of both shape and position of the hand. The feature vector for the n^{th} frame in the video sequence is selected to be:

$$\underline{f}(n) = [A(n)\, C_x(n)\, C_y(n)\, C_{xx}(n)\, C_{xy}(n)\, C_{yy}(n)]^T$$

where $C_{xx}(n)$, $C_{xy}(n)$ and $C_{yy}(n)$ are the second order spatial moments of the hand region detected as a skin-colored blob, $C_x(n)$ and $C_y(n)$ are the first order moments and $A(n)$ is the zeroth order moment. $A(n)$ is the area of the skin-colored blob whereas $C_x(n)$ and $C_y(n)$ are the coordinates of the centroid of the skin colored blob. Thus $A(n)$ relates to the magnification of the imaging system, $C_x(n)$ and $C_y(n)$ capture the hand movement. The second-order moments capture the temporal change in the appearance of the palm. Thus our feature set comprises of both positional and appearance information. If required, one can use higher order spatial moments, but we found the above set of moments to possess sufficient discrimination for the given vocabulary.

Fig. 2. Examples of few gestures from our vocabulary. Rows above describe frames of RIGHT, UP, CCW and CLOSER gestures.

To have meaningful values from one video sequence to another, we scale the first and second order moments by the width of the hand region during the start position, and the zeroth order moment by the initial area, i.e. $A(0) = 1$. This gives invariance to magnification due to the distance from the camera at which the gesture is performed, and also with respect to changes in hand size from person to person. Thus we have,

$$A(n) = m_{00}(n)/m_{00}(0); C_x(n) = m_{10}(n)/m_{01}(0)$$

$$C_y(n) = m_{01}(n)/m_{01}(0); C_{xx}(n) = m_{20}(n)/m_{01}(0)$$

$$C_{xy}(n) = m_{11}(n)/m_{01}(0); C_{yy}(n) = m_{02}(n)/m_{01}(0)$$

where $m_{pq}(n)$ is the pq^{th} moment of the hand region obtained by tracking in the n^{th} frame.

The raw data obtained by tracking may be noisy; hence we filter the data using a median filter of length 3. To remove the effect of normalizing the trajectories in time between [0:1] and to further smoothen the trajectories, we filter the data using an averaging filter of length 7. Figure 3(a) shows the raw features as obtained from the tracker for the RIGHT gesture, clearly depicting their noisy nature. The result of filtering this using a median and an averaging filter is given in Figure 3(b). Similarly filtered feature trajectories of CCW and UP gestures performed by the same user have been shown in Figure 3(c) and Figure 3(d), respectively.

4. PRINCIPAL COMPONENT ANALYSIS

In the past researchers have used principal component analysis (PCA) for classification, specifically, gesture recognition [10] [12]. Our task is to classify the gesture into one of the eight classes based on six feature trajectories $\underline{f}(n)$. Feature trajectories of same gesture performed by different users were found to be correlated. One approach is to capture the variation in the collection of feature trajectories

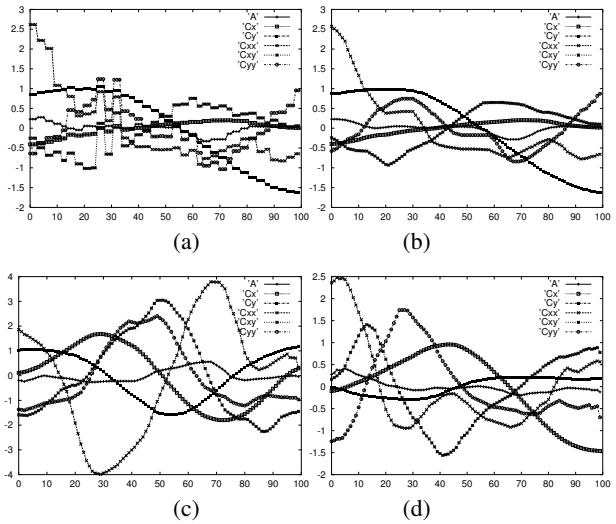

Fig. 3. Example feature trajectories. (a) Unfiltered features for **Right** gesture. (b), (c) & (d) Filtered features for RIGHT, CCW & UP gesture.

(training data set) and use this information to compare unknown feature trajectories. This is done by finding the principal components of the distribution of feature trajectories, or the eigenvectors of the covariance matrix of the set of feature trajectories. The eigenvectors are ordered, each one accounting for a different amount of variation among the feature trajectories. We call these eigenvectors, *eigentrajectories*. Each feature trajectory can be approximated using only those eigentrajectories corresponding to largest eigenvalues as they account for the maximum variance within the feature space.

Calculating Eigentrajectories

Each feature vector of dimension N (N=100 in our study) is a point in N-dimensional space. Features for gestures belonging to the same class tend to cluster in a small region in this N-dimensional space, and thus can be described by a relatively low dimensional subspace. The main idea of PCA is to find vectors that best describe the distribution of feature trajectories in the feature space. Now if λ_1, λ_2,...,λ_M are the training feature trajectories corresponding to any one of the moments for any one class of gesture, Δ being their mean trajectory, then the covariance matrix R is given by VV^T, where

$$V = [\lambda_1 - \Delta, \lambda_2 - \Delta, ..., \lambda_M - \Delta]. \quad (1)$$

Thus, we get a covariance matrix of dimension $N \times N$. This implies that N eigenvectors are to be determined. We consider the eigenvectors v_i (corresponding to eigenvalues η_i) of V^TV,

$$V^TVv_i = \eta_i v_i \quad (2)$$

Premultiplying eqn(2) by V, we get:

$$VV^TVv_i = \eta_i Vv_i$$

We see that Vv_i are the eigenvectors of the covariance matrix $R = VV^T$. The computations are reduced from the order of N to the number of training feature trajectories M (Typical values of M are around 5).

5. GESTURE RECOGNITION USING EIGENTRAJECTORIES

The M eigentrajectories computed using the above method span the subspace describing the corresponding feature trajectory. We compute and store the eigentrajectories for all the features from all the classes along with the corresponding mean trajectories.

Any new gesture is then characterized by a set of six feature trajectories $\lambda_A, \lambda_{Cx}, \lambda_{Cy}, \lambda_{Cxx}, \lambda_{Cxy}$ & λ_{Cyy} corresponding to $A, C_x, C_y, C_{xx}, C_{xy}$ & C_{yy}. We reconstruct each trajectory by projecting them on the corresponding eigentrajectories of each gesture,

$$\hat{\lambda}_i = \sum_{j=1}^{M} < \lambda_i, \lambda_{ij} > \lambda_{ij} \quad (3)$$

where $i = A, Cx, Cy, Cxx, Cxy, Cyy$ and $< \lambda_i, \lambda_{ij} >$ is the projection of λ_i on λ_{ij}, the j^{th} eigentrajectory of feature i. In actual practice, we may not use all the M eigen trajectories to reconstruct λ_i. We reconstruct $\hat{\lambda}_i$ by projection on all gesture classes and evaluate the reconstruction error,

$$E = \sum_{i=A,Cx,Cy}^{Cxx,Cxy,Cyy} \| \lambda_i - \hat{\lambda}_i \|^2 \quad (4)$$

for each gesture class. The gesture is classified as belonging to a class for which the reconstruction error is minimum. Eqs(3) and (4) can be combined and the error can be computed using:

$$E = \sum_{i=A,Cx,Cy}^{Cxx,Cxy,Cyy} \{< \lambda_i, \lambda_i > - \sum_{j=1}^{M} (< \lambda_i, \lambda_{ij} >)^2\} \quad (5)$$

6. EXPERIMENTAL RESULTS

The gesture recognition scheme was tested on a data set of 112 gestures performed by different users. Each gesture sequence is of different length varying from 40 to 80 frames depending on the time taken to perform the gestures. We normalize the duration of all gestures to 100 frames. The gestures were captured using a low-cost camera at a framerate of 12 frames per second, in a natural office environment with a cluttered background. Out of the 112 gestures,

No.	Gestures	Instances	Testing	Recognised
1	LEFT	16	11	10
2	RIGHT	13	8	8
3	UP	11	6	6
4	DOWN	16	11	11
5	CCW	13	8	8
6	CW	13	8	8
7	AWAY	15	10	7
8	CLOSER	15	10	10
Total		112	72	68

Table 1. Summary of experimental results

only 40 were used for training, 5 from each of the 8 gesture classes. The remaining 72 gestures were used to test the proposed scheme. Initially, all $(M = 5)$ the eigentrajectories were used for reconstruction and the recognition rate was 94.4%. Then we reduced the number of eigentrajectories used for reconstruction gradually to 2. In this case too the results remained the same, but it was observed that the error in reconstruction given in eqn(5) increased. However, the cumputation is faster as we need to project only on two eigentrajectories. Table (1) summarizes the results.

7. CONCLUSIONS

In this paper, a simple technique for gesture recognition using the data available from a hand tracker developed earlier has been proposed. We use PCA of feature trajectory to recognize unknown gestures. Our recognition technique is unaffected by the rate at which gestures are performed. The computations reduced due to reduction in dimension of feature space make the scheme computationally feasible. We observe very good accuracy even for a small number of principal components. The reason for this is that we have considered variation in position and appearance of the hand in motion. Currently we are working on extending the gesture vocabulary.

8. REFERENCES

[1] R. Sharma, V. Pavlovic, and T. S. Huang, "Visual interpretation of hand gestures for human-computer interaction: A review," in *IEEE Trans. on PAMI*, 1997, vol. 19, pp. 677–695.

[2] J. Davis and M. Shah, "Visual gesture recognition," in *IEE Proc. - Vision, Image, Signal Processing*, April 1994, vol. 141, pp. 101–105.

[3] J. Weaver, T. Starner, and A. Pentland, "Real-time american sign language recognition using desk and wearable computer based video," in *IEEE Trans. on PAMI*, Dec 1998, pp. 1371–1375.

[4] C. Vogler and D. Metaxas, "Asl recognition based on a coupling between hmms and 3d motion analysis," in *Proc. of IEEE ICCV*, Mumbai, India, 1998.

[5] M. Yeasin and S. Chaudhuri, "Visual understanding of dynamic hand gestures," in *Pattern Recognition*, 2000, vol. 33, pp. 1805–1817.

[6] J. Davis and A. Bobick, "The representation and recognition of action using temporal templates," in *Proc. of IEEE CVPR*, San Juan, Puerto Rico, 1977.

[7] Hu, "Visual pattern recognition by moment invariants," in *IRE Trans. Info. Theory*, 1962, vol. IT-8(2).

[8] M. H. Yang and N. Ahuja, "Recognizing hand gesture using motion trajectories," in *Proc. of IEEE CVPR*, Colorado, June 1999.

[9] S. McKenna and S. Gong, "Gesture recognition for visually mediated interaction using probablictic event trajectories," in *Proc. of BMVC*, Southampton, UK, September 1998.

[10] J. Martin and J.L. Crowley, "An appearance-based approach to gesture-recognition," in *Proc. of ICIAP*, Florence, Italy, September 1997.

[11] H. Ohno and M. Yamamoto, "Gesture recognition using character recognition technique on 2d eigenspace," in *Proc. of IEEE ICCV*, Greece, 1999.

[12] D. Hall Martin and J. L. Crowley, "Statistical recognition of parameter trajectories for hand gestures and face expressions," in *Proc. of ECCV*, Germany, June 1998.

[13] A. F. Bobick and A. D. Wilson, "A state-based approach to the representation and recognition of gesture," in *IEEE Trans. on PAMI*, Dec 1997, pp. 1325–1337.

[14] M. J. Black and A. D. Jepson, "Recognizing temporal trajectories using the condensation algorithm," in *Proc. of IEEE AFGR*, Nara, Japan, April 1998.

[15] Tushar Agrawal and Subhasis Chuadhuri, "Gesture recognition using motion histogram," in *Proc. of NCC*, Chennai, India, Jan 2003.

[16] Matthew Turk and Alex Pentland, "Eigenfaces for recognition," in *Journal of Cognitive Neuroscience*, 1991, vol. 3(1):pp, pp. 71–86.

[17] James Mammen and S. Chaudhuri, "A model based technique for palm detection," in *Proc. of NCC*, Kanpur, India, Jan 2001, pp. 315–319.

[18] S. Chaudhuri James Mammen and Tushar Agrawal, "Simultaneous tracking of both hands by estimation of erroneous observations," in *Proc. of BMVC*, Manchester, Sep 2001.