

VISUAL HAND MOVEMENTS FOR MACHINE CONTROL

Sanjay Kumar, Dinesh Kant Kumar, Arun Sharma

School of Electrical and Computer Engineering RMIT University GPO Box 2476 Melbourne, VIC 3001, Australia

Keywords: Hand gestures classification, motion based representation, Computer Vision, K-NN method Mahalanobis distance.

Abstract: A new technique for automated classification of human hand gestures for robotics and computer control applications is presented. It uses view-based approach for representation, and statistical technique for classification. This approach uses a cumulative image-difference technique where the time between the sequences of images is implicitly captured in the representation of action. This results in the construction of Temporal History Templates (THTs). These THT's are used to compute the 7 Hu image moments that are invariant to scale, rotation and translation. The recognition criterion is established using K-nearest neighbor (K-NN) mahalanobis distance. The preliminary experiments show that such a system can classify human hand gestures with a classification accuracy of 92%. Our research has been carried on in the robotics framework. The overall goal of our research is to test for accuracy of the recognition of hand gestures using this computationally inexpensive way of dimensionality-reduced representation of gestures for its suitability for robotics.

1 INTRODUCTION

Research in improving human-computer interaction systems has resulted in the development of a variety of systems that have applications in fields such as virtual reality, telemedicine and computer games. An important part of these systems is the input module that is devoted to recognize the command by the human operator.

Dynamic hand actions are the basis of hand gestures and play a very important role in the interactions between people. But the interaction of people with computers is based static events such as a key press. Information contained in the dynamic gesture is lost and this reduces the scope of the control of the machine. To improve human interaction with computer machines and for robotics applications, it is desirable for machines to extract more information from human hand movement.

Systems reported in literature may be classified into two broad categories; (i) Requiring the user to wear or hold some device (ii) Using video data. Most of the systems reported in literature are invasive and require the use of gloves (Akita,1984 , Baudel, 1993), reflectors (N Ma,2001), or electrodes (Poole,2002). In the recent past, video data based

non-invasive techniques to identify human activity have been reported. Fong et al presented a virtual joystick technique based on static gestures to drive remote vehicle (Terence Fong and Charles Baur,1996), in which hand motions are tracked with a color and stereovision system. The system depends on the static gesture and the interface is not user friendly.

Baudel et al developed a system called 'Charade' to control remote objects using free-hand gestures (Lafon, July, 1996) . Using Charade, a speaker giving a presentation can control remote computer display with free-hand gestures while still using gestures for communicating with audience. The system has the problem of use of data glove and accuracy of classification is a major concern. Another technique reported uses an 'elastic graph', a conductive sensor, to classify hand postures against complex backgrounds in gray-scale images (Jochen Triesch and Christoph von der Malsburg, 2001). But this technique is invasive and unable to cope with are the large variations in the shape of hands referring to the same posture.

Moy (Moy,1999) and Bretzner (Lars Bretzner and Lenman) have proposed visual interpretation of 2D dynamic hand gestures in complex environments. It is used for humans to communicate

and interact with a pet robot (Moy, July 18-22 1999) and control home appliances (Lars Bretzner and Lenman, 1997). But these techniques require hand segmentation and feature extraction making the system not transportable. The inability of the techniques to reliably track the hand, and dependency on the background lighting and gesture positioning makes these unsuitable for HCI applications.

Laptev et al used particle filtering and hierarchy object models to track multi-state hand models (Ivan Laptev and Tony Lindeberg, 2001). But the approach is very prone to variable background conditions.

A complex, neural network based users and a mobile robot (Boehme, Sep. 17~19, 1997) interface proposed by Triesch is based on 4 cue modules sensitive to skin color, facial structure, structure of a head-shoulder-contour and motion. But this system lacks robustness to the environmental constraints.

An intuition based system to provide naturalness suitable for two hands has been developed by Caroline et al for computer supported product design (Caroline Hummels, Sep. 17~19, 1997). This interface supports the perceptual-motor skills and is task-specific.

The previous techniques for hand gesture identification have been generally too intrusive, unreliable, or computationally complex (Akita, 1984, Baudel, 1993) (Davis, April 1994) (Hinton, Jan 1993) (Sturman, Jan, 1994). These methods are user dependent and lack naturalness.

The present work is view-based approach for the representation and classification of pre-defined gestures using characteristics of the fine motion of hand-gestures from particular view direction using video data. The technique is based on the work of Bobick and Davis (Davis, November 1998) (Aaron F. Bobick, 2001) and the authors (Arun Sharma, WITSP'2002) and uses of Temporal History Template (THT). This research has combined the use of THT with the image moment technique proposed by Hu (Hu, 1962). The recognition criterion is achieved by using K-NN nearest neighbor technique using Mahalanobis distance. The technique is computationally simple and results demonstrate robustness.

2 THEORY

The technique presented in this paper is based on the spatio-temporal templates of hand movements for recognition. This "THT" is a single static image integrated over time, is very distinctive for short duration actions and is considered to be spatio-

temporal templates of hand movements. The motion features of the THT are computed using geometrical Hu moments and classified using Mahalanobis statistical distance.

Videos of pre-defined hand actions of a group of people are recorded. Temporal History Templates corresponding to the different actions are generated and stored in a database. From the THT of the various hand gestures, global shape descriptors are extracted corresponding to each hand movement. Statistical distances are computed and used for classification of the test recordings. During the recognition phase, the hand gestures of the user are recorded, THT generated and Hu moments computed. The action is identified using K-NN (K-Nearest Neighbor) classifier, "Mahalanobis distance". Details of each of these are described below.

2.1 Temporal History Templates

This paper reports the use of statistical properties of the geometric moments of THT to identify hand movements. The representation of THT is a view-based approach of hand action representation. The technique is based on collapsing the hand motion over time to generate a static image from the image sequence. This resulting static image is representative of the whole sequence of video frames of the hand movement. Normalisation of the image is used to overcome the difference in speed of the action. This technique is very suitable for short duration, non-repetitive, medium velocity movements making it suitable for real-time computer interface application (Arun Sharma, WITSP'2002.).

2.1.1 Motion Image Estimation

For this work a simple temporal difference of frame technique (DOF) has been adopted (Aaron F. Bobick, 2001). The approach of temporal differencing makes use of pixel difference between two or three consecutive frames in an image sequence to extract motion information (Aaron F. Bobick, 2001). The DOF technique subtracts the pixel intensities from each subsequent frame in the image sequence, thereby removing static elements in the images. Based on research reported in literature, it can be stated that the actions and messages can be recognized by description of the appearance of motion (Davis, November 1998) (Sanjay Kumar) (Pentland, July 1997) (Starner, 1995) (Aaron F. Bobick, 2001) (Arun Sharma, WITSP'2002.) (Sanjay Kumar) without reference to underlying static images, or a full geometric reconstruction of the

moving hand (Little, November 1995). It can also be argued that the static images produced using THT based on the DOF represent features of temporally localized motion (Davis, November 1998) (Aaron F. Bobick, 2001) (Arun Sharma, WITSP'2002.) (Sanjay Kumar 2001). This process can be represented mathematically as follows

Let $I(x, y, n)$ be an image sequence

&

DOF be $D(x, y, n) = |I(x, y, n) - I(x, y, n-1)|$

Where $I(x, y, n)$ is the intensity of each pixel at location x, y in the n th frame and $D(x, y, n)$, is the difference of consecutive frames representing regions of motion.

$B(x, y, n)$ is the binarisation of image difference over a threshold of Γ

$$B(x, y, n) = \begin{cases} 1 & \text{if } D(x, y, n) > \Gamma \\ 0 & \text{otherwise} \end{cases}$$

Putting a ramp multiplier to represent time results in the THT. In a THT H_N , pixel intensity is a function of the temporal history of motion at that point. The result is a scalar-valued image where more recently moving pixels are brighter (Davis, November 1998) (Aaron F. Bobick, 2001) (Arun Sharma, WITSP'2002.) (Sanjay Kumar, 2002).

$$\begin{aligned} \text{THT}(H_N(x, y)) \\ = \text{Max} \bigcup_{n=1}^N I \quad B(x, y, n) * n \end{aligned}$$

Where N represents the duration of the time window used to capture the motion.

2.2 Feature Extraction

Hand gestures produce grey scale THT with global features and with variations due to the rotation and change in scale. Thus it is important to extract global features of the static image that are scale, translation and rotation invariant. Hu moments are invariant to scale, rotation and translation are based on the geometrical normalised centralised moments of the image (Hu, 1962).

The definition of the zero-th order geometric moment, m_{00} , of the image $f(x, y)$ is

$$m_{00} = \sum_{x=1}^N \sum_{y=1}^M f(x, y)$$

The two first order moments, $\{m_{10}, m_{01}\}$ identify the centre of mass (light intensity) of the object. This defines a unique location that may be used as a reference point to describe the position of the object within the field of view. The coordinates of the centre of mass can be defined through moments as shown below

$$\bar{x} = m_{10}/m_{00}, \quad \bar{y} = m_{01}/m_{00}$$

According to uniqueness theory of moments for a digital image of size (N, M) the $(p+q)$ th order moments m_{pq} are calculated for $p, q = 0, 1, 2, \dots$

$$m_{pq} = 1/NM \sum_{x=1}^N \sum_{y=1}^M f(x, y) x^p y^q$$

The centralized moments, μ_{pq} , of the image provides the translation invariance and can be calculated as showed below:

$$\mu_{pq} = 1/NM \sum_{x=1}^N \sum_{y=1}^M f(x, y) (x - \bar{x})^p (y - \bar{y})^q$$

$f(x, y)$ is intensity function of the gray scale image.

2.2.1 Feature Classification

Based on the above, identification of the hand gestures requires classification of the seven dimensional Hu moments of the THT. This can be achieved using statistical approaches or by artificial neural networks. Among the supervised training statistical approaches, Bayesian technique is most common. But this requires assumption of appropriate probability densities, which could be a matter of concern for seven dimensional feature

space. The other technique is classification by Hidden Markov Model (HMM). The main drawback of HMM is its probabilistic approach and their relatively modest discriminative power for classification. Among the statistical technique, the K-nearest neighbor (KNN) technique is commonly used. Mahalanobis distance is an efficient measure for KNN classification. The advantage of the method is its computational simplicity.

2.3 Mahalanobis Distance

The MAHALANOBIS distance is a statistical technique of determining the "similarity" of a set of values from an "unknown" sample to a set of values measured from a collection of "known" samples. It is computed by the equation below:

$$r^2 \equiv (f - k_x)' C^{-1} (f - k_x)$$

where r is the Mahalanobis distance from the feature vector f to the mean vector k_x , and C is the covariance matrix for f .

3 METHOD

To test the technique, experiments were conducted where five subjects were asked to make five pre-defined hand gestures; the Move "Clasp" gesture (MC), the Move "Right" gesture (MR), the Move "Left" gesture (ML), Move "hold" gesture (MH) and Move "Grab" gesture (MG) -Figure 1. Each hand action was performed and recorded for duration of 3 second at frame rate of 30 frames/sec. The movement was recorded using a video camera at a distance of 1.2 meters from the hand and a with a window size of 0.09 sq meters. The video data was stored as true color (AVI files) with an array size of 120*160 for each frame. All the computing was done using Image analysis package in Matlab 6.1.

These AVI files were transformed to eight-bit gray scale images (0-255 levels). The duration of the movement was determined from manually located delimiters, and this determined the number of frames for each gesture and thus the duration of integration of the DOF to generate the THT. To take care of the variation in speed, the intensity image for THT was normalized between [0.... 1] before computing the image moments. From the THT representation of each action 7-Hu moments were computed. There are total 150 actions samples used for classification purposes into five classes. The data was divided into subsets of training data, validation, and test subsets. One fourth of the data was used for

the validation set, one-fourth for the test set, and one half for the training set. During the recognition the user is asked to perform the test hand gesture, from the test hand gesture THT is generated and features are extracted to be compared with the pre-stored features of the various THT's using K-NN (K-Nearest Neighbor) classifier, "Mahalanobis distance". The feature vectors whose Mahalanobis distances are minimum are classified as the members of the class.

4 RESULTS AND DISCUSSION

The results of the testing show that with the system described can classify the five gesture classes with 92 % accuracy (Table 1). This accuracy can be attributed to the invariance to variations such as rotation, scale and translation of Hu Moments, and also due to the better discriminating ability of DOF technique. Reasons for inaccuracy in discrimination can be attributed to the image differencing technique being sensitive to pixels revisited while performing the hand action.

Table 1 shows the results of classification

Class	No of Actions	Predicted Membership of Classes					Accuracy (%)
		MC	MR	ML	MH	MG	
MC	30	27	-	1	-	1	90 %
MR	30	1	29	1	-	-	97 %
ML	30	-	1	26	2	-	87 %
MH	30	2	-	1	28	1	94 %
MG	30	-	-	1	-	28	93 %

5 CONCLUSIONS

This paper reports the testing of a new technique for identifying hand actions using video data using Hu-moments of the THT as features and with K-NN nearest neighbor for classification. Temporal integration of the video sequences of the hand movements removes the static content from the video sequences to generate THTs of the hand movement. Experiments suggest that THTs of different classes present distinctive 2-D motion patterns where each pixel describes the function of temporal history of motion in that sequence. The scale, translation and rotation invariant features have been used for discrimination of the THT for classification. On the basis of the experimental results it can be concluded that the THT based

method can be successfully used for computer interaction and for controlling robots. The system is fast and accurate when compared to other attempts for machine based hand movement gesture recognition. In the current form, the limitation of the method is its suitability to short duration hand movements only. The next step is to test the accuracy on large database, long duration movements and the sensitivity analysis of the method. Future work also investigates the effect of noise on the classification accuracy.

REFERENCES

- Aaron F. Bobick, J. W. D. (2001) The Recognition Of Human Movements Using Temporal Templates. IEEE - Pattern Analysis and Machine Intelligence, 23 No 3, 257-267.
- Akita, K. Pattern recognition, 1984, Image sequence analysis of real world human motion. 17 (No.1), 73-83.
- Arun Sharma, D. K. K., Sanjay Kumar, Neil McLachlan (WITSP'2002.). Representation and Classification of Human Movement Using Temporal Templates and Statistical Measure of Similarity Workshop On Internet Telecommunications and Signal Processing. 2002. Wollongong, Sydney Australia.
- Baudel, T., Beaudouin-Lafon, M (1993) Charade: remote control of objects using free hand gestures CACM, 1993, 28 -35.
- Boehme, H.-J. (Sep. 17~19, 1997) International Gesture workshop Bielefeld, Germany, 2 Neural Architecture for Gesture-based Human-Machine-Interaction". 13~232.
- Caroline Hummels, G. S., and Kees Overbeeke (Sep. 17~19,1997) An Intuitive two-hands gestural interface for computer support product design International Gesture workshop Bielefeld, Germany, 197~208.
- Davis, J. a. A. B. (November 1998) Virtual PAT: a virtual personal aerobics trainer Proc. Perceptual User Interfaces.
- Davis, J. S., M (April 1994) *Visual gesture recognition*. Vision, Image and Signal Processing IEE Proceedings *Vision, Image and Signal Processing IEE Proceedings*, 141(Issue: 2), 101 -106.
- Hinton, S. S. F. a. G. E. (Jan 1993) "Glove-talk: a neural network interface between a data-glove and a speech synthesizer IEEE Trans. on Neural Networks,, vol-4, 2--8.
- Hu (1962) Visual Pattern Recognition By Moment Invariants IEEE - Pattern Transaction On Information Theory, 8(2), 179-187.
- Ivan Laptev and Tony Lindeberg (2001). Jochen Triesch and Christoph von der Malsburg. Tracking of multistage hand models using particle filtering and a hierarchy of multi-scale image feature
- Lafon, T. B. a. M. B. (July, 1996) Remote control of objects using free-hand gestures", Communications of ACM Communications of ACM, 3, 36(7), 28~35.
- Lars Bretzner, I. L., Tiny Lindeberg, Soren and Lenman, a. Y. S. (Jan 1997) A prototype system for computer vision based human computer interaction Proc of the workshop on imaging and neural networks Little, J., and J. Boyd (November 1995) Describing motion for recognition International Symposium on Computer Vision,, 235-240.
- Moy, M. C. (July 18-22 1999) Gesture-based Interaction with a Pet Robot Proceedings of 6th National Conference on Artificial Intelligence and 11th Conference on Innovative Applications of Artificial Intelligence Orlando, Florida, USA, 628-633.
- N Ma, D. K. K., N D Pah ANZIIS 2001 Classification of Hand Direction using Multi-Channel Electromyography by Neural Networks PR106.
- Pentland, I. E. a. A. (July 1997) Coding, Analysis, Interpretation, and Recognition of Facial Expressions IEEE Trans. Pattern Analysis and Machine Intelligence, 19, no. 7, 757-763.
- Poole, E. D. K. K. IEEE EMBS 2002, Classification of EOG for Human Computer Interface USA.
- Sanjay Kumar, A. S., Dinesh Kant Kumar, Neil McLachlan (2002) Classification of Visual Hand Gestures Using Difference of Frames Proc. of the Int. Conf. on Imaging Science and Technology, Las Vegas, Nevada, USA , CISST'02. 2002. Las Vegas, USA: (CSREA Press, 2002).
- Starner, T. P., A. (1995) Visual Recognition of American Sign Language Using Hidden Markov Models Proc. Intl Workshop on Automated Face and Gesture Recognition Zurich, 1995.
- Sturman, D. J. Z., D (Jan, 1994) *A survey of glove-based input* ".14 (Issue: 1), 30-39.
- Terence Fong, F. C., (June 1996) Sebastien Grange, and Charles Baur Novel interfaces for remote driving: gesture, haptic and PDA", 11th IEEE conference on Image Analysis








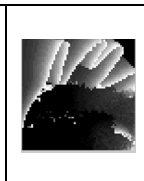



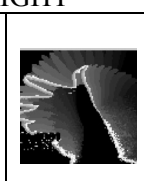
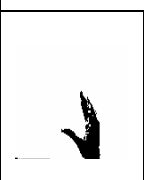
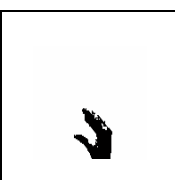




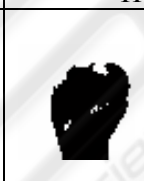

First Frame	Intermediate Frame	End Frame	THT
			
Move "CLASP"		Move "CLASP"	
			
Move "Right"		Move Identifier "RIGHT"	
			
Move "LEFT"		Move Identifier "LEFT"	
			
Move "HOLD"		Move Identifier "HOLD"	
			
Move "GRAB"		Move Identifier "GRAB"	

Figure 1: Visual Hand Movement Sequences and THTs