2004 IEEE Workshop on Machine Learning for Signal Processing

# SENSITIVITY ANALYSIS OF HAND MOVEMENT CLASSIFICATION TECHNIQUE USING MOTION TEMPLATES

Sanjay Kumar, Dinesh K Kumar, Arun Sharma
School of Electrical and Computer Engineering
RMIT University, GPO Box 2476V Melbourne, Victoria, Australia 3001
Phone:00-61-3-99253025 Fax-00-61-3-99252007
Email:s2003383@student.rmit.edu.au

## ABSTRACT

**This paper presents the sensitivity analysis of a new technique for automated classification of human hand gestures based on Hu moments for robotics applications. 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 THTs 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%. This research has been conducted for medical and 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 medical and robotic applications.**

## 1. INTRODUCTION

There is a well-recognized need to improve the human computer interaction systems that will give the user a natural way of controlling machines. Research 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 [1] [2], reflectors [3], or electrodes [4]. 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 [5], 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. The previous techniques for hand gesture identification have been generally too intrusive, unreliable, or computationally complex [1] [2] [6] [7] [8]. These methods are user dependent and lack naturalness.

The authors have reported a 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 [9]. The technique uses Temporal History Template (THT) along with the image moment technique proposed by Hu [10]. The recognition criterion is achieved by using K-NN nearest neighbor technique using Mahalanobis distance. The technique is computationally very simple and can easily be supported for real time applications.

For this technique to be used in the real world, it is important that it has to be robust and lay people are able to use it without the need for extensive training. It is important that the technique is user independent and various factors such as position and angle of the camera, lighting conditions and background have minimal effect on the ability of the system to identify the hand actions accurately. With this aim, the features for the THT need to be rotation, translation and scale invariant and noise resilient. This research reports the use of Hu moments that are rotation and translation invariant. This paper reports the experimental results to study the sensitivity analysis on the technique of these factors.

0-7803-8608-6/04/$20.00 ©2004 IEEE

## 2.THEORY

The technique developed by the authors is based on the spatio-temporal templates of hand movements for recognition [9]. "THT" is a single static gray scale image integrated over time and is very distinctive for short duration actions and thus suitable for hand movement identification.

Hand gestures produce grey scale THT with global features. The image intensity is dependent on the change of pixel intensity at each point between frames caused by the motion of the object assuming the lighting to remain constant. Any change in the position of the object with respect to the camera results in the translation, rotation or scaling of the THT. 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 normalized centralized moments of the image [10].

Variation in the speed of motion results in a variation of the number of frames representing the action and thus the intensity of the THT. Small inter experiment variations in the action results in small variation of the THT and can be modelled as addition of noise on the image. Hence, it is important to identify the THT with the hand action using features that are noise resilient and image intensity invariant. For this purpose the authors have normalized the intensity of the image to overcome the variation in the speed of motion.

It is also important to have the classifier that is able to successfully separate the various features in a complex multi-dimensional space. Mahalanobis statistical distance is a technique for classifying multi-dimensional features and are very reliable in complex, multi-dimensional fuzzy space [11]. These techniques are described below.

### 2.1 Temporal History Templates

For this work a simple temporal difference of frame technique (DOF) has been adopted [12]. The approach of temporal differencing makes use of pixel difference between two or three consecutive frames in [13] an image sequence to extract motion information [12]. 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 [11] [12] [13] [14] [15] [16] without reference to underlying static images, or a full geometric reconstruction of the moving hand [17]. It can also be argued that the static images produced using THT based on the DOF represent features of temporally localized motion [11] [12] [14] [13]. 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.

$\boldsymbol{B(x, y, n)}$ is the binarisation of image difference over a threshold of $\Gamma$

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

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 [11] [12] [14] [13]. THT $(H_N(x, y))$ is:

THT $(H_N\ (x, y))$

$$=\mathbf{Max}\left\{\bigcup_{n=1}^{N-1} B(x,y,n)*n\right.$$

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 [10].

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

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

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

$$\overline{X} = m_{10}/m_{00}$$

$$\overline{Y} = m_{01}/m_{00}$$

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

$$m_{pq}=1/NM\sum_{x=1}^{N}\sum_{y=1}^{M} f(x, y)\, x^p y^q \tag{2}$$

The centralized moments, $\mu_{pq}$, of the image provides the translation invariance and can be calculated and showed below.

$$\mu_{pq}=1/NM\sum_{x=1}^{N}\sum_{y=1}^{M} f(x, y)\,(x-\bar{x})^p (y-\bar{y})^q \tag{3}$$

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

### 2.2.1 Feature Classification

Identification of the hand gestures requires classification of the seven dimensional Hu moments of the THT to the hand gestures. This project reports the use of K-nearest neighbor (K-NN) technique with Mahalanobis distance as a measure of the distance. This has been used because this does not require any assumption of the statistical property of the data and is very efficient for multidimensional data. Another advantage is its computational simplicity.

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:

$$r^2 \equiv (\mathbf{f} - \mathbf{k}_x)' \ \mathbf{C}^{-1} (\mathbf{f} - \mathbf{k}_x)$$

where $r$ is the Mahalanobis distance from the feature vector $\mathbf{f}$ to the mean vector $\mathbf{k}_x$ and $\mathbf{C}$ is the covariance matrix for $\mathbf{f}$.

## 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) , Move "Grab" gesture (MG) -Figure 4. Each hand action was performed and recorded for duration of 3 second at frame rate of 30 frames/sec. 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. These AVI files were transformed to eight-bit grey scale images (0-255 levels). The experiments were repeated on five subjects. The conditions were changed and experiments repeated to determine the sensitivity of the technique to variation in lighting, change in rotation in image plane, and change in scale etc. These are detailed below:

| S.No | Move Description | Move Identifier |
|---|---|---|
| 1 | MOVE "CLASP" | MC |
| 2 | MOVE "RIGHT" | MR |
| 3 | MOVE " LEFT" | ML |
| 4 | MOVE "HOLD" | MH |
| 5 | MOVE "GRAB" | MG |

**Table 1 Movement Identifier Codes**

### 3.1 Sensitivity to Variable Lighting

The changing lighting and illumination conditions can play an important role in the THT. Experiments were conducted to determine the effect of variation of lighting on the accuracy of identification of the hand gestures using this method. For this purpose, experiments with natural (window) and varying (using florescent lamps) lighting conditions were conducted for each subject. The system was tested by using different lighting conditions between the training data and testing data. The results are tabulated in the section 4.

### 3.2. Sensitivity to Variable Background

One of the challenges faced by video based gesture recognition systems has been the sensitivity to the background conditions. When the background is complex and varying, the hand may be seen against lighter background in some places and seen against darker background in other places in the same frame and between frames, making it difficult to track the hand accurately and identify the action. The problem is also compounded because the boundary of the hand may be undetectable where hand and background have the same grey-level values. For these reasons the experiments were conducted to test the sensitivity of the THT based approach against different background conditions; uniform light colored background, uniform dark background and complex background.

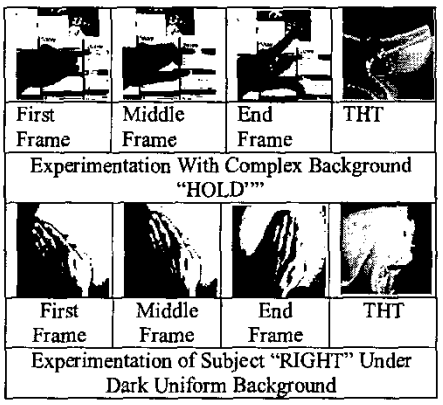


**Figure 1** Sensitivity Experiments with Complex Background and Dark Background

### 3.3 Sensitivity to Scale

The sensitivity to the change in scale due to change of distance between the camera and hand on the ability to accurately identify the hand action was determined experimentally. Experimentations were conducted for predefined hand movements at three different scales Figure 2, inter-scale relationship being an extension of 10 %.

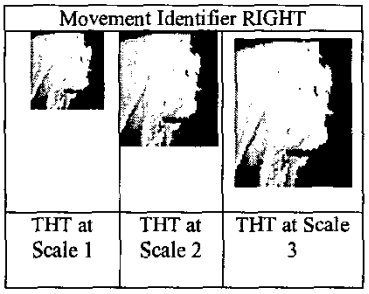


**Figure 2** Sensitivity Experiments for Scale

### 3.4 Sensitivity to Rotation

The sensitivity of this system to the view/rotation, experiments were conducted and predefined hand movements were captured at different angles from $180^{\circ}$ to $90^{\circ}$ at an interval of $15^{\circ}$ Figure 3.

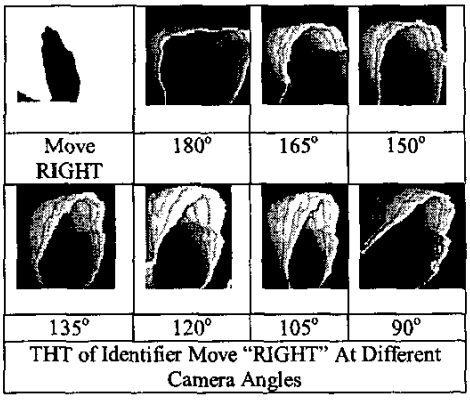


**Figure 3** Sensitivity Experiments at Different angles in image plane

### 3.5 Data Analysis

The recorded AVI files were transformed to eight-bit grey scale images (0-255 levels). To take care of the variation in speed, the intensity image for THT was normalized between [0.... 1]. Image moments were computed for the normalized THT images. From the image moments, the 7 Hu moments were computed for each of the hand gestures for each of the experiments. A total of 150 actions samples were used. 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 testing, the user was asked to perform a hand gesture similar to one of the five recorded during training and a corresponding THT was generated and Hu moments computed. Using Mahalanobis distance, these were used to identify the gesture. The number of correctly identified gestures as a percentage of the total number of tests was the indicator of the accuracy of the system. This was computed for each of the conditions.

## 4. RESULTS AND DISCUSSION

The results of the accuracy of the system for hand actions with a static lighting, light colored background and with fixed camera angle and position are tabulated in Table 2. Table 3 describes the results of sensitivity analysis of the method. From the results it is observed that the system described can classify the five gesture classes with 92 % accuracy when the lighting is fixed and the background is simple and static. The results during sensitivity analysis indicate that the THT based method of recognition is sensitive to variable lighting but is invariant to variable background, scale, and to the view and angles or rotation.

| Class | No of Actions | Predicted Membership of Classes | | | | | |
|---|---|---|---|---|---|---|---|
| | | MC | MR | ML | MH | MG | Accuracy (%) |
| 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 %** |

**Table 2** Confusion Matrix for classification data

The inaccuracy of the system when there is fixed lighting, background and relative position of the camera and hand is due to the noise indicating inter-experiment differences. The ability of the system to identify the action despite variations in static lighting, background and change in the relative position of the camera and the hand is due to the choice of THT and Hu moments to represent the action. THT is invariant to the variations in the background while Hu moments are invariant to variations such as rotation, scale and

translation. The invariance to the background is attributed to the THT based method and DOF technique, which integrates the motion while removing any static contents from the movement. The large error when the lighting is varying is attributed to the THT identifying change in lighting (due to external lighting) as an indicator of motion resulting in a large error.

| Sensitivity Analysis Results | | | | | | |
|---|---|---|---|---|---|---|
| Results of Variable lighting conditions | | | | | | |
| Movement Identifier | MC | MR | ML | MH | MG | Average Accuracy (%) |
| Natural Light | **90** | **97** | **87** | **94** | **93** | **92.2** |
| Variable Lighting | **57** | **48** | **51** | **46** | **42** | **48.8** |
| Results of Variable Background conditions | | | | | | |
| Light Uniform Background | **90** | **97** | **87** | **94** | **93** | **92.2** |
| Dark Uniform Background | **91** | **89** | **88** | **95** | **90** | **90.6** |
| Complex Background | **84** | **86** | **93** | **91** | **98** | **90.4** |
| Results of Variable Scale | | | | | | |
| Scale1 | **90** | **88** | **85** | **97** | **87** | **89.4** |
| Scale 2 | **96** | **90** | **89** | **94** | **82** | **90.2** |
| Scale3 | **85** | **93** | **87** | **96** | **88** | **89.8** |
| Results of Variable Angle | | | | | | |
| 90° | **86** | **87** | **95** | **90** | **97** | **91** |
| 135° | **83** | **88** | **87** | **95** | **91** | **88.8** |
| 180° | **93** | **96** | **96** | **90** | **88** | **92.6** |

**Table 3** Results of Sensitivity Analysis

## 5. CONCLUSION

This paper has tested the sensitivity of a new method of hand action identification using THT along with Hu- moments and K-NN nearest neighbor for classification. The low level representation of the action collapses the temporal structure of the motion from the video sequences of the hand movements removing any static content from the video sequences. The scale, translation and rotation invariant features have been used for discrimination of the THT for classification. On the basis of the preliminary experimental results it can be concluded that the THT based method can be successfully used to identify pre-defined hand actions when the lighting conditions are invariant. The sensitivity analysis results indicate that the technique is scale, background, rotation, and translation invariant but is sensitive to light. Future work is required to investigate the effect of noise on the classification accuracy.

## 6.REFERENCES


[1] Akita, K., *Image sequence analysis of real world human motion.* Pattern recognition. **17 (No.1)**: p. 73-83., 1984

[2] Baudel, T., Beaudouin-Lafon, M, *"Charade: remote control of objects using free hand gestures".* CACM: p. 28 -35, 1993

[3] N Ma, D.K.K., N D Pah, *Classification of Hand Direction using Multi-Channel Electromyography by Neural Networks.* ANZIIS 2001

[4] Poole, E.D.K.K., *Classification of EOG for Human Computer Interface.* IEEE EMBS 2002, USA.

[5] Terence Fong, F.C., Sebastien Grange, and Charles Baur, *"Novel interfaces for remote driving: gesture, haptic and PDA", http://vrai-group.epfl.ch/papers/SPIE00-tf.pdf-*

[6]. Davis, J.S., M, *"Visual gesture recognition.* Vision, Image and Signal Processing IEEE Proceedings. **141(Issue: 2)**: p. 101 -106,April 1994

[7] Hinton, S.S.F.a.G.E., *"Glove-talk: a neural network interface between a data-glove and a speech synthesiser."* IEEE Trans. on Neural Networks. **Vol-4**: p. 2--8, Jan 1993

[8] Sturman, D.J.Z., D, *"A survey of glove-based input "*. IEEE Computer Graphics and Applications, **14(Issue: 1)**: p. 30-39. Jan 1994.

[9] Sanjay Kumar, A.S., Dinesh Kant Kumar, Neil McLachlan, *"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).

[10] Hu, *Visual Pattern Recognition By Moment Invariants*. IEEE - Pattern Transaction On Information Theory,**8(2)**: p. 179-187, 1962.

[11] Davis, J.a.A.B., *"Virtual PAT: a virtual personal aerobics trainer"*. Proc. Perceptual User Interfaces, November 1998.

[12] Aaron F. Bobick, J.W.D., *The Recognition Of Human Movements Using Temporal Templates*. IEEE - Pattern Analysis and Machine Intelligence**23 No 3**: p. 257-267. , 2001.

[13] Arun.Sharma, D.K.K., Sanjay Kumar, Neil McLachlan, *Classification of Human Actions using Temporal Templates, Histograms And Orthogonal Moments*. SCI"2003 Proceedings Of The 7th World Multiconference Conference On Systemics Cybernetics and Informatics Orlando, Florida, USA, 24th-27th July 2003.

[14] Arun Sharma, D.K.K., Sanjay Kumar, Neil McLachlan, *"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, WITSP'2002.

[15] Pentland, I.E.a.A., *"Coding, Analysis, Interpretation, and Recognition of Facial Expressions "*. IEEE Trans. Pattern Analysis and Machine Intelligence, **19, no. 7**: p. 757-763. July 1997

[16] Starner, T.P., A., *"Visual Recognition of American Sign Language Using Hidden Markov Models"*. Proc. Intl Workshop on Automated Face and Gesture Recognition Zurich, 1995., 1995.

[17] Little, J., and J. Boyd, *" Describing motion for recognition"*. International Symposium on Computer Vision,, November 1995: p. 235-240.

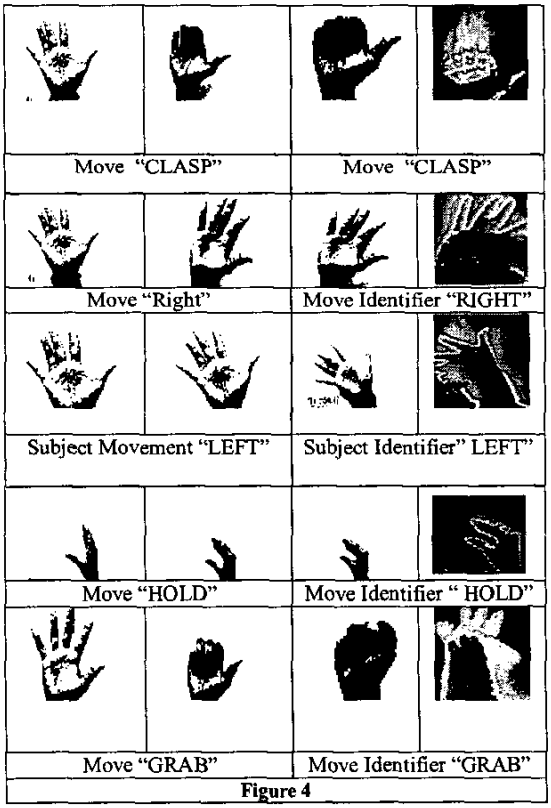


**Figure 4**