Classification Of Human Gestures Using Temporal Templates And Artificial Neural Networks
Arun Sharma, Dinesh K Kumar, Sanjay Kumar, Neil McLachlan
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
e-mail: s9900960@student.rmit.edu.au

Abstract
The aim of this research is to determine techniques for machines to recognize human actions using image-based spatio-temporal approach. This paper presents work that determines the inter and intra subject reliability in the ability of automated recognition of pre-defined actions. This work is motivated by the previous research in appearance-based motion recognition of human actions. The approach reported in this paper uses a cumulative image-difference technique that results in the construction of Motion History Image (MHI) where a single image represents a sequence of frames covering motion related to a complex human gesture. Based on the Hu image invariant moments of MHI, templates for a pre-defined actions are generated and are tested by classifying similar actions by different subjects. Supervised backpropagation neural networks have been used for classifying the data.

Index Terms—Human action classification, motion based representation, spatio-temporal, Artificial Neural Networks, computer vision.

I. INTRODUCTION
Human actions are complex sequence of movements and may be task oriented such as eating or walking, or intended to communicate. These movements may also contain more information than the result of the task itself (if task oriented) or the intended message (if communicative). They may contain information about the action, the quality of action, the actor and the environment. However, to date, most of the information extracted by machines from human movement has been from static events such as a key press. To improve machine surveillance, human interaction with machines and for helping disabled people, it is desirable for machines to extract more information from human movement.

Past attempts to recognize dynamic actions by machines reported in literature require intrusive devices that limit the scope of their applications to situations where people specifically intend to communicate with computers [1-9]. Other major early works involved the use of Moving Light Display (MLD) on subjects in a darkened room[10, 11] and Structure from Motion (SFM) techniques where a 3 dimensional model of the person is reconstructed to recognise the action [12-16]. While MLD was a useful experiment and demonstrated human perceptual abilities from motion information alone, the MLD technique is very intrusive and it lack naturalness and robustness. Systems using SFM [9, 17] techniques are more computationally expensive, and also assume the body to be a 'rigid object', thereby greatly limiting the applications.

This paper reports the research conducted to identify simple pre-defined human gestures of subjects from video data using computationally simple techniques that do not require the use of intrusive devices, which are not user dependent and not lacking in naturalness. This is the first step towards developing applications where movement data may be interpreted by machines in normal living and working environments to identify the gestures or the subject. Examples of such applications may include identification of messages, actions, status and identity of individuals operating specific equipment or inhabiting a specific environment. This will take the capability of machines into the 'understanding people' domain.

This research has incorporated the techniques proposed by Bobick and Davis [18, 19] Black and Yakoob [20] Romer Rosales & Stan Sclaroff [21]. The most relevant approach related to our work, as in [18] [19], is to recognize movement directly from sequence of images. Davis [18] used a view based technique to represent and recognize actions by building Motion History Images (MHI) and Motion Energy Images (MEI). The recognition of temporal templates in their technique used moment based features for representation and nearest neighbour for recognition against examples of pre-defined movements.

This research uses MHI combined with Hu invariant image moment technique [22] to represent gross body movements and classification [19] using artificial neural networks (ANN) applied to moment-based features [23]. This paper reports the intra and inter subject reliability study of this technique in the classification of the pre-defined simple gross human body actions.

II. THEORY
AN OVERVIEW AND EXPLANATION OF THE APPROACH

Based on research reported in literature, it can be stated that actions and messages can be recognized by description of the appearance of motion [10, 11, 19, 24-31] without reference to underlying static images, and without a full geometric reconstruction of the moving part [20]. It can also be argued that the static images produced using MHI based on the Difference of Frames (DOF) can represent features of temporally localized motion [32]. The advantage of this technique is the computational simplicity. This paper reports the efforts to determine the reliability of this technique.
II (a) REPRESENTATION OF MOTION:

The DOF data from video recorded human gestures are analysed by temporal integration of the frames covering the period of the distinct movement. The delimiters for the start and stop of the movement are added manually in the sequence. The temporal history of the movement in MHI is inserted into the data by multiplication of the intensity of each frame with a linear ramp representing time. The MHI grey scale images are then generated by temporal integration. This eliminates the need for time sequence coding. The process can be mathematically represented by the following:

Let \( I(x, y, t) \) be the intensity of each pixel at location \( x, y \) at time \( t \) in an image sequence and \( D(x, y, t) \), the difference of consecutive frames. Then:

\[
D(x, y, t) = I(x, y, t) - I(x, y, t-1)
\]

\( \text{MHI} (H_r(x, y, t)) \) is:

\[
H_r(x, y, t) = \begin{cases} 
\tau & \text{if } D(x, y, t) = 1 \\
\max(0, H_r(x, y, t-1) - 1) & \text{Otherwise}
\end{cases}
\]

where \( \tau \) represents the duration of the time window used to capture the motion. The pixel intensity \( H_r \) is a function of motion history at that point and the result is a scalar-value image where intensity is function of the recency of motion [19]. The MEI can be generated by thresholding MHI above zero.

II (b) FEATURE EXTRACTION:

Given an arbitrary intensity function \( f(x, y) \), the set of moments \( m_{pq} \) (equation 2) have a property of fundamental importance: they uniquely determine, and are uniquely determined by, the function \( f(x, y) \) and are sufficient to reconstruct the original function \( f(x, y) \) [37, 38]. For an appearance based, view sensitive approach, it is desirable to have matching technique that is invariant to the imaging situation [18]. It needs to be invariant to translation (space in the image where object is represented), rotation and scaling. The moments defined by equation 1 are not ideal for MHI description since they are not invariant to translation, rotation and scale. To overcome this difficulty, modified seven Hu moments [22] (equation 4) have been used.

Seven Hu's equations are based on the uniqueness theory of moments. According to uniqueness theory of moments for a digital image of size \( (N, M) \) the \((p+q)\)th order moments \( m_{pq} \) are calculated

\[
m_{pq} = \sum_{x=1}^{N} \sum_{y=1}^{M} f(x, y) x^p y^q \]

The central moments of a digital image are inherently translation independent,

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

where

\[
\bar{x} = m_{10}/m_{00}
\]

and

\[
\bar{y} = m_{01}/m_{00}
\]

and

\[
\mu_{00} = m_{00} = 1
\]

\( \mu_2 = 0 \)

\( \mu_{12} = 0 \)

\( \mu_{20} = x^2 \)

\( \mu_{11} = 2y x \)

\( \mu_{02} = y^2 \)

\( \mu_{30} = 3x_3 - 3x \bar{x}^2 + 24 \mu_1 x \)

\( \mu_{13} = 3x_2 y - 2x_1 y^2 + 8 \mu_2 y \)

\( \mu_{21} = 3x_1 y + 2x^2 y + 8 \mu_3 y \)

\( \mu_{03} = 3x_0 y^2 + 2 \mu_4 y^3 \)

To achieve invariance with respect to orientation and scale, first normalize for scale defining \( n_{pq} = \mu_{pq}/(\mu_{00})^{(p+q)/2} \)

Where

\( \gamma = (p+q)/2 + 1 \) and \( p+q \geq 2 \)

The first seven Hu moments are defined as Hu's seven moment functions below utilize the central moments of a digital silhouette or boundary image, but are also rotation independent.

\[
M_1 = (n_{02}, n_{00})
\]

\[
M_2 = (n_{20}, n_{00})^2 + 4n_{02}n_{00}
\]
The calculation of the feature vectors involves using normalized central moments (equation 3). These moments are reputed to be invariant to affine transformations, such as translation. Using these moments, the invariant nature of the feature vectors can be taken a step further by applying a set of moment functions called, "Hu functions" which result in translation, rotation and scale invariant features. These invariants are based on moments up to third order.

The advantage of moment methods is that they are mathematically concise and for the intensity image of MHI, reflect not only the shape but also the density distribution within it. Thus, the descriptors of the Motion templates of human gestures generated by MHI have been calculated using the Hu derived set of seven functions that make use of the central moments of an image.

When classifying human actions, there is an uncertainty regarding the speed of motion and thus the number of frames that represent the action. To overcome this problem, this research has used normalising of the MHI's such that these lie within the same intensity range. The intensity range for faster moves is relatively expanded and that of slower moves is contracted. This is explained in greater detail in Section III (b).

II ( c) CLASSIFICATION:
The reliability of classification using MHI for gesture representation and Hu moments for features can be accomplished by supervised learning techniques. This can be achieved using statistical approaches or by artificial neural networks. Among the supervised training statistical approaches, Bayesian technique is most common. Bayesian method requires assumption of appropriate probability densities. Due to high dimensionality of feature space, this is extremely difficult. Another statistical technique, the K-nearest neighbour (K-NN) a non-parametric statistical classifier, does not assume any probability distribution, is most suitable but it is very slow for large datasets. An alternative approach, which circumvents the determination of probability densities and does not suffer from the "curse of dimensionality" is based on the idea of discriminant functions. This simple discriminant can be generalised by transforming the linear combination with a non-linear function (called an activation function) that leads to concepts such as perceptron [39]. One reason for the strong interest in multilayer perceptron system, such as back propagation, is that they are able to select useful input features from high-dimensional input vectors and are capable of implementing more complex partitioning of feature space when there are sufficient samples representing the data.

This paper reports the use of artificial neural networks (ANN) to the image moments of the human gestures performed by people for classification. This work aims to determine conclusive answers for the above mentioned problem using multilayer perceptron of neural network against examples of given motions already learned.

MULTILAYER PERCEPTRON (MLP) SYSTEM

The multilayer perceptron of neural network, once trained by an appropriate set of training vectors, can assign any input pattern to one of several output classes. Thus the input pattern is an action represented by Hu moments, and the output is the class to which the action belongs. One weakness of the standard ANN architecture is its inability to deal with temporally varying delimiters such as in human movement video data. One way to handle this weakness is the use of recurrent neural networks [40] that employ input layer of the ANN augmented by hidden context units that give feedback to a hidden layer, based on the previous contents of the hidden layer. This gives the network a 'memory' of past events [41] and thus a moving window is achieved. But this technique increases the dimensionality of the problem, of concern when the number of frames representing the data may be large. But by the use of MHI representing human actions, having removed the need for temporal domain of the data (number of frames), ordinary feed forward (FF) neural network with back propagation learning algorithm (BPN) can be used to classify actions.

A two-layer feed forward perceptron (MLP) system with one hidden layer has been adopted (Figure 5). Four pre-defined human actions are classified by a two-layer perceptron. The input to the MLP consists of seven Hu moments of the MHI of the actions to be classified. The two-layer perceptron has seven inputs, 50 hidden nodes and 4 (number of classes) output nodes.

During the training of the network, the target pattern belonging to the n-th class has a desired output with 1 in the n-th output node and 0 in the others. The network uses sigmoid as threshold function and gradient descent with momentum and adaptive learning as training algorithm.

II. METHOD

The experiments were designed to determine the intra and inter subject reliability of classification of actions when using MHI as representation of motion, Hu moments to represent the MHI and ANN as a classifier. Five subjects performed four pre-defined actions and these were repeated ten times each.

The human actions approximate to the Normal Distribution in that it has low variation probability values with large numbers. According to the theory of the Normal distribution, a smaller standard deviation, σ, gives better confidence when the PDF is more “tighter”.

\[
\begin{align*}
M_3 &= (n_{10} - 3n_{12})^2 + (3n_{21} - n_{00})^2 \\
M_4 &= (n_{10} + n_{12})^2 + (n_{02} + n_{20})^2 \\
M_5 &= (n_{10} - 3n_{12}, (n_{10} + n_{12})^2) + (3n_{21} - n_{00}, (n_{21} + n_{02})^2) + (3n_{02} + n_{20})^2 \\
M_6 &= (n_{10} - n_{10}, (n_{10} + n_{12})^2, (n_{21} + n_{02})^2) + (3n_{02} + n_{20})^2 \\
M_7 &= (n_{10} - 3n_{12}, (n_{10} + n_{12})^2, (3n_{21} - n_{00}, (n_{21} + n_{02})^2) - (n_{10} - 3n_{12}, (n_{10} + n_{12})^2, (3n_{21} - n_{00}, (n_{21} + n_{02})^2))^{-1}
\end{align*}
\]
When \( N \) readings are taken from each test, the later confidence analysis will be based on the mean \((X')\) and range \((R)\). If \( N \) is less than 10, then:
\[
X' = \frac{\sum(X_i)}{N}, \quad \text{i from 1 to N, } X_i \text{ are reading values,}
\]
\[
R = \text{Max}(X_i) - \text{Min}(X_i), \quad \text{i from 1 to N.}
\]
\[
\sigma_r = \frac{R}{D_2/\sqrt{N}},
\]
where \( D_2 \) is a factor used to calculate upper control limit (UCL) and lower control limit (LCL).

This relation shows that more the readings taken, the smaller the value of \( \sigma_r \). But after the number of readings exceed 6, the curve becomes “flattened”. This means the increase of the number of readings has less effect on decreasing of the value of \( \sigma_r \). So, the optimal balance between confidence level and working load is that each test should be approximately six readings. This is also true for statistical averaging of the data.

For each of the experiments, the following constraints were maintained during motion capture:

I. Related to Movements
- SCMO (Stationary Camera Moving Object)
- Presence of Single Subject In Scene.
- Constant Window Size and View Angle
- The Subject Remains Inside the Workspace.
- Movements are parallel to the camera plane.
- Duration is small and same pixels are not revisited by motion.

II. Related to Environment and Subject:
- Consistent background and illumination.
- Static Background
- Known Start pose
- No restriction on colour and tightness of clothes.
- Subjects of different height, weight, age and both genders.

III (a) EXPERIMENTATION

Experiments were conducted where the subject was asked to make four pre-defined gestures: 'Arm Wave - Full and Half', 'Standing to Sitting' and 'Lifting Tea Cup from Table to Lips' (Figures 1 - 4). The movements were recorded using a video camera in 2-D vision-space at 2 meters normal to subject and the window size was an area of 6 sq meters. Each subject was asked to repeat the experiment ten times and five volunteers were used for collecting the data.

The steps of training and classification of gestures are as follows:
1. Compute the Motion History images for 40 gestures of each of the four classes.
2. Generate Seven Hu moments as the features for the training and testing gestures.
3. Input 7 features of each training gesture to two layer perceptron for training until the network converges.
4. Classify the testing gestures.

III (b) SIGNAL ANALYSIS

The video data was stored on the PC and each frame was stored as an array of size 120 X 160 and the recordings were in true colour (AVI files). All the computing was done using Image analysis package in Matlab 6.1.

These AVI files were later transformed to eight-bit grey scale image (0-255 levels) for further processing. The duration of the movement was determined from the 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 MHI. To take care of variation in speed, the intensity image for MHI is normalised between [0.1] before computing moments. For each of the image representing motion so produced, the seven Hu image moments [36] were computed which were used to train the neural network and classify the gestures.

III. RESULTS AND DISCUSSION.

MHI's were generated for each of the four gestures (figure 1-figure 4). The results of the testing showed that with the use of Artificial Neural Network as classifier and Hu Moments as the feature space, the classification accuracy is significant and is the likely range of the true value.

The confidence of the technique is evident from the inter-subject and intra-subject classification accuracies shown in Tables 1 and 2. Table 1 is result of 10 test samples from each class comprising of two samples of each of the five subjects. Table 2 is the result for individual subject for 10 samples of each class. The subjects were of different height, weight and gender. The results clearly show that the method is invariant to all such factors. Inclusion of extra constrains like tight clothes will improve confidence level at the cost of naturalness.

RESULT OF CLASSIFICATION:

<table>
<thead>
<tr>
<th>Gesture</th>
<th>S</th>
<th>D</th>
<th>LRH</th>
<th>LRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy %</td>
<td>100</td>
<td>85</td>
<td>85</td>
<td>83</td>
</tr>
</tbody>
</table>

Table 1. Inter – Subject percent accuracy of classification

<table>
<thead>
<tr>
<th>Actions</th>
<th>S</th>
<th>D</th>
<th>LRH</th>
<th>LRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy %</td>
<td>100</td>
<td>83</td>
<td>81</td>
<td>80</td>
</tr>
</tbody>
</table>

Table 2. Intra – Subject percent accuracy of classification

(S – Sit; D – Drink; LRH – Lateral Arm Down Half; LRF – Lateral Arm Raise Full)

Reasons for inaccuracy in discrimination can be attributed to the fact that the image differencing technique is very sensitive to secondary motion of the body part (eg. Loose clothes) which may not essentially be a part of the gesture; and also constant intensity pixels in consecutive frames do not capture the motion in spite of subject movement. The explanation for 100% accuracy in the gesture “Sitting on a chair” is due to the fact that the whole body movement is included and the loose clothes do not add to noise. Background Subtraction is likely to take care of these problems.
The results show that the method looks at the appearance of what motion looks like and is independent of the person performing the gesture. Two persons performing the same gesture will generate very similar MHI and MEIs. Besides being rotation, scale and translation invariant, it is person invariant also. Insensitivity to subject makes this technique very robust to classify gestures.

It should also be mentioned that even though the process of temporal integration greatly reduces the data and the processing time is greatly reduced, the computation of image moments was found to be a relatively slow process.

The other issue of concern is the need for delimiters. For practical applications of this technique, it would be necessary to overcome that difficulty.

CONCLUSIONS:

This paper reports the technique for classification of dynamic human actions. It uses MHI to represent the movement making it computationally very efficient. Hu moments have been used to represent the MHI and this being rotation, scale and translation invariant, is highly reliable in classification of actions. It has been experimentally determined that this technique is intra and inter subject reliable and actions can be classified irrespective of the actor.

One of the obvious disadvantages is that it cannot be used to distinguish between two subjects performing the same move. An addition of more vector components containing additional information about the gesture may be able to overcome this drawback. Also, longer duration moves where the motion revisits the same pixel again, is not picked up by MHI. Extension of the method to include discrimination among subjects and to take care of longer moves needs to be worked upon.

Another matter of concern is the high computational expense for computing Hu moments. But the authors are now attempting to use faster algorithms for this purpose [42].

References:


23. Huang, K.-Y., Neural Network for robust pattern recognition in real seismic data. National Chiao Tung University, Hsinchu, Taiwan, 1999.


27. M.J. Black and Y. Yakkob. Tracking and recognizing rigid and non-rigid facial motion using local parametric models of image motion. in Internation Conference of Computer Vision.


