
Available from Deakin Research Online:

http://hdl.handle.net/10536/DRO/DU:30044650

Reproduced with the kind permissions of the copyright owner.

Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Copyright : 2002, IEEE
Hierarchical Recognition of Intentional Human Gestures for Sports Video Annotation

Graeme S. Chambers, Svetha Venkatesh, Geoff A.W. West, and Hung H. Bui
School of Computing, Curtin University of Technology, Perth, Western Australia
E-mail: {chambegs,svetha,geoff,buihh}@cs.curtin.edu.au

Abstract

We present a novel technique for the recognition of complex human gestures for video annotation using accelerometers and the hidden Markov model. Our extension to the standard hidden Markov model allows us to consider gestures at different levels of abstraction through a hierarchy of hidden states. Accelerometers in the form of wrist bands are attached to humans performing intentional gestures, such as umpires in sports. Video annotation is then performed by populating the video with time stamps indicating significant events, where a particular gesture occurs. The novelty of the technique lies in the development of a probabilistic hierarchical framework for complex gesture recognition and the use of accelerometers to extract gestures and significant events for video annotation.

1. Introduction

Significant research attention has recently been focused on the characterisation of video sequences. Characteristics of video can be partly determined by detection of important events within the video. Events aimed to be detected should be significant in terms of the video being processed. For example, we may wish to automatically determine action scenes in films, or, of concern here, recognise gestures of humans indicative of significant video events. Detection of such gestures is an unconstrained, difficult problem when only video data is considered. Our approach is to recognise intentional human gestures by augmenting video with data obtained from accelerometers worn by one or more humans such as sporting officials. This is useful as gestures made by sporting officials correspond to specific events in the game, enabling automatic generation of highlights.

We refer to a gesture as a specific, intentional action by a human in which part of the body is moved in a predefined way indicating a specific event. We consider a gesture as a stochastic process which exists at multiple levels in a hierarchy — simple, discrete movements correspond to gestures at one level, and combinations of these movements as gestures at a higher level.

The novelty of our approach is threefold: First, it enables the recognition of complex gestures using a multi-layer hidden Markov model (HMM), an extension of the HMM for modeling complex gestures as a hierarchy of sub-gestures. This essentially enables a rich sequence of sub-gestures to be recognised, without the need of learning separate HMMs for each sequence. The hierarchy also allows for segmentation of sub-gestures allowing learning of new sequences on-line. Second, the aim of the gesture recognition is to populate a sports video with umpire gestures, each of which correspond to specific events. This multi-modal approach allows for robust and semantically meaningful annotation. Lastly, the use of accelerometers to derive semantic events and aid video annotation is novel.

2. Background

Gesture recognition has been explored in both image and sensor domains. Sensor-based techniques have the advantage that computationally intense calculations are not required for accurate movement information as measurements are provided directly by the sensors. Many researchers working in the image-domain argue that cameras are unobtrusive compared to the often cumbersome sensor devices. Sensor-based techniques, however, have another advantage in that they can be used in much less constrained domains and are not reliant on lighting conditions or camera calibration. The advent of micro machining has enabled dramatic reduction in the size of some sensors. Our accelerometers for example, are very lightweight and can be easily fitted to bracelets or wrist bands.

Previous work in vision based gesture recognition has concentrated on recognition of individual gestures in constrained environments. Hand sign language recognition [9] and learning of single T'ai Chi movements in virtual environments [3] are two examples. Sensor based recognition systems have also been limited in their scope. Many systems developed aim to recognise hand sign language using
complex gloves, such as the Acceleration Sensing Glove [2] and VPL DataGlove [7], containing many sensor devices.

The recognition performance of these systems is heavily reliant on the complexity of gestures considered for recognition. Measurements of human gestures are inherently noisy as there is no clear distinction between when a sub-gesture stops and another sub-gesture begins. This poor definition of sub-gesture boundaries and the high degree of variability in gestures lends recognition to a hidden Markov model approach.

The hidden Markov model, as the name suggests, requires the Markov condition to be satisfied. The condition, however, is inappropriate for gesture recognition as a sequence of gestures does not depend on only the previous time step, but also the previous sub-gesture or gesture, and the contextual meaning of the entire gesture. For this reason, many researchers have proposed extensions of the model. The parametric hidden Markov model [8], for example, introduces a parameterisation between output probabilities of the hidden states, attempting to model gestures which exhibit spatio–temporal correspondence, such as "this big", or "over there". Another extension to the standard HMM is the Coupled [4] and Factorial HMMs [6]. In the case of Coupled HMMs, two standard HMMs are coupled together such that their transitions from one state to another are linked. Although these models provide suitable extensions to the standard HMM for gesture recognition, such as increasing the ability of the system to model temporal influences, they overlook modeling the hierarchical nature of human gestures.

3. Extension to the Hidden Markov Model

Our framework for recognition extends the standard hidden Markov model by introducing a hierarchy of interacting hidden states. This model can be considered as a special type of Abstract HMM [5] and also relates to other extensions such as Coupled/Factorial HMMs where the hidden chain is replaced by a number of interacting hidden chains. A significant distinction between the AHMM and Coupled/Factorial HMMs is the way in which the hidden chains interact. Each chain in the AHMM only interacts with chains directly below and above it. This type of interaction allows the focus to be on the dynamics of temporal abstraction among the chains rather than just the correlation between them at the same time interval. The abstraction also allows us to consider gestures and sub-gestures individually and in combination. We can consider simple, individual movements as gestures at one level, and combinations of the simple movements at a higher level. Further, we can consider sequences of combination movements at an even higher level. Figure 1(a) illustrates the way in which gestures and sub-gestures interact. The lowest level consists of a number of standard HMMs, each corresponding to an individual movement. The lowest level is entirely independent of the layer above it in terms of learning. Higher layers represent the order of sub-gestures which form a gesture. The lower layer independence allows us to train individual sub-gestures with varying numbers of hidden states and levels of complexity. The only commonality required in the lower levels is the type and dimensionality of observations. Viewing the model "unrolled" over time illustrates the hierarchical interaction between the layers, Figure 1(b). If there is only two layers, such as the case in Figure 1, the second (upper) layer corresponds to the sub-gesture $U_t$ at the current time $t$ and the first layer as the state $S_t$ inside the sub-gesture model for $U_t$. The second layer can be viewed as the dynamic changes between the sub-gesture models. The recognition performance of lower layers is obviously crucial when considering upper layers, as the overall performance is highly influenced by the lowest level.

3.1. Training

As each sub-gesture is independent, each can be trained individually. If there are $k = 1..M$ sub-gestures, the result of training is a set of state transition probabilities $A_k$, observation probabilities $B_k$, and initial state distributions $\pi_k$. The order in which these sub-gestures appear is specified by a Markov chain $D$ at the upper layer. This models the evolution of the sub-gesture from one time point to the next.

$$D(k, l) = P(U_{t+1} = l | U_t = k)$$  \hspace{1cm} (1)

Learning the second layer is trivial. Using the ground truth of complete gestures, the number of occurrences of each sub-gesture, and the number of transitions from one sub-gesture to another can be calculated, then normalised.

3.2. Hierarchy Construction

If there are two layers, such as the case in Figure 1, we can construct a joint distribution of the probability of a sub-
gesture and the probability of a hidden state within a sub-gesture.

\[ P(U_{t+1}, S_{t+1} | U_t, S_t) \]  

(2)

The two layer case then becomes a HMM with \( N \times M \) hidden states. Each state within the joint distribution HMM has a clear semantic meaning, representing a state within a particular sub-gesture.

A HMM is specified by three probability distributions. The transition probabilities of the joint HMM are formed using Equation (3) and the state transition probabilities of each of the sub-gesture models. We first require the probability of moving from one sub-gesture \( k \) to another sub-gesture \( l \). This is simply the Markov chain already calculated.

\[ P(U_{t+1} = l | U_t = k, S_t = i) = D(k, l) \]  

(3)

Given the sub-gesture, we now require the probability of moving from one state within a sub-gesture to a state in possibly another sub-gesture, given Equation (4).

\[ P(S_{t+1} = j | U_{t+1} = l, U_t = k, S_t = i) = \begin{cases} A_l(i, j) & \text{if } l = k \\ \pi_l (j) & \text{otherwise} \end{cases} \]  

(4)

If there is a switch from one sub-gesture to another, the previous state of the previous sub-gesture has no influence on the first state of the next sub-gesture, as reflected in the hierarchy of Figure 1(a). The observation probabilities of the HMM, \( P(O(t)|S_t = i, U_t = k) \), are simply the observation probabilities of each of the sub-gesture models, \( B_k(O(t)|S_t = i) \). The final distribution required is the initial state probabilities. Using an estimated initial probability of sub-gestures at the first time step \( (D_0) \), the initial state probabilities are given by Equation (5).

\[ P(U_0 = k, S_0 = i) = \pi_k (i) D_0(k) \]  

(5)

Using this representation allows us to recognize any sequence of combinations of sub-gestures, with each combination requiring only a Markov chain to be specified. The standard HMM would require training of a new model for each new combination sequence. The re-use of sub-gesture models in our technique provides a convenient and efficient method for learning new sequences. It also avoids the problem of increasing state space since training of each level is independent.

4. Recognition of Gestures

Figure 2 shows the architecture of the overall system. As the time-stamped accelerometer data is analyzed and the gesture recognized, the video is annotated with the event corresponding to the specific gesture. The accelerometer devices used [1] measure acceleration in two orthogonal directions. Mounted on wrist bands along the plane of the wrist, the directions correspond to the lower arm and the thumb. The accelerometers are connected to a PC serial port, via a prototype board, with a sampling rate of 96Hz. Only one of the accelerometers is currently used as the gestures considered are performed with only one arm. The features extracted from the accelerometer data include the zero crossing rate of the first and second derivatives, root mean square, and mean using a window size of 32 samples with 8 samples overlap.

5. Experimental Results

Our dataset consists of several Kung Fu martial art movements acted out by an instructor in a simulated training video. Each of the individual movements are modeled as sub-gestures and combinations of the movements as gestures. Figure 3(a), for example, shows the gesture "cuts" consisting of the sub-gestures "wood", "grass", "throat", and "side". The hierarchical modeling of our approach can be seen from Figure 3(b). The figure displays the probability of each sub-gesture within the gesture "cuts" for an unseen sequence. The figure shows that the most likely sequence of sub-gestures is "wood", "grass", "throat", and "side". The recognized sequence demonstrates the ability to accurately segment a gesture into its comprising sub-gestures and annotate the video with time stamps accordingly.

Two other complete gestures are considered for testing our approach. The "elbows" gesture corresponds to elbow movements forward then to the side, and the "punch blocks" gesture is comprised of a forward punch, a block, then a sideways punch. A training set consisting of 10 examples of each of the nine sub-gestures is used to learn the HMM parameters for each model in the lower layer. The sub-gestures are left-right models with two hidden states.

The test set consists of 30 unseen sequences of the complete gestures, 10 instances of each gesture described. Table 1 lists the resulting confusion matrix. The table shows that our system can robustly differentiate between gestures.

<table>
<thead>
<tr>
<th>Cuts</th>
<th>Elbows</th>
<th>Punch Blocks</th>
<th>Classified As</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>Cuts</td>
</tr>
<tr>
<td>1</td>
<td>9</td>
<td>0</td>
<td>Elbows</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>10</td>
<td>Punch Blocks</td>
</tr>
</tbody>
</table>

Table 1. Gesture Confusion Matrix
6. Conclusions

A technique for the recognition of complex human gestures has been presented. The novelty of the work is in the way in which gestures are recognised and the application to populating sports video with time stamped events corresponding to gestures of sporting officials. Our framework for recognition allows us to represent gestures and sub-gestures at different levels of a hierarchy. The hierarchy also allows us to recognise a rich sequence of sub-gestures without the need of constructing separate hidden Markov models for each sequence.

References


