Automatic Labeling of Sports Video Using Umpire Gesture Recognition

Graeme S. Chambers, Svetha Venkatesh, and Geoff A.W. West

Department of Computing, Curtin University of Technology, Perth, Western Australia
{chambegs,svetha,geoff}@cs.curtin.edu.au

Abstract. We present results on an extension to our approach for automatic sports video annotation. Sports video is augmented with accelerometer data from wrist bands worn by umpires in the game. We solve the problem of automatic segmentation and robust gesture classification using a hierarchical hidden Markov model in conjunction with a filler model. The hierarchical model allows us to consider gestures at different levels of abstraction and the filler model allows us to handle extraneous umpire movements. Results are presented for labeling video for a game of Cricket.

1 Introduction

Content characterisation and labeling of video sequences have become significant research areas with the goal of automatically describing arbitrary video. Extracting high level semantics from video data is a difficult problem. To begin annotation, there must be knowledge of the domain of video to be processed and some limitations imposed on the types of scenes that can be analysed. Rather than increasingly confining the domain of video to be processed, we consider the problem on a broader scope by introducing other sensors. If other sensor data is available, the difficulties associated with image processing can be avoided. In particular, we have accelerometer sensors in the form of wrist bands which can be worn by key actors in the video. In sports video, for example, umpires in the game can wear the sensors and have their movements recorded and analysed throughout the game.

Sports officials perform many gestures which are indicative of what is going on in the game. Their gestures can provide something meaningful about a player, a team, or the entire game. If the gestures of these officials are able to be recognised, meaningful information can be derived. We refer to a gesture as an intentional action whereby part of the body is moved in a predefined way to indicate a specific event. Detecting these events enables automatic generation of highlights and more importantly, rich, contextual labeling of video. To solve this problem we need to address the issues of segmenting continuous gesture data and performing robust gesture classification.

The novelty of the work presented here is the way in which we segment and classify candidate gestures from the continuous stream. We propose the use of the Hierarchical Hidden Markov Model (HHMM) in conjunction with a filler model for segmenting and classifying gestures at differing levels of detail. The HHMM allows us to consider gestures as sequences of sub-gestures, possibly reusing the sub-gesture parts for gestures.

A. Fred et al. (Eds.): SSPR&SPR 2004, LNCS 3138, pp. 859–867, 2004.

The filler model allows us to potentially ignore unknown movements and automatically segment and classify gestures simultaneously from a continuous stream.

2 Background

The types of gestures we are interested in detecting are intentional movements by officials which indicate something about the game. In many sports, umpires move to an area where other officials can see them, perform the gesture that represents the event in the game, then return to officiating. Obviously these gestures can provide a first level for semantic labeling of the accompanying video. In the area of sports video, several attempts have been made at meaningful labeling, including specific sports [1, 2] and automatic generation of highlights [3]. These however, do not provide suitable reusable frameworks for recognising events in various types of sports. All assume specific knowledge of the domain and have heuristics for the sport being processed.

Gesture recognition from video sequences has also been limited in scope, primarily focused on individual gestures in constrained environments. Hand sign language recognition [4] and learning of T'ai Chi movements [5] are two such examples. Attempting to recognise real world gestures places much more demand on the image segmentation techniques and generally results in dramatically increased failure rates. Lighting conditions and occlusion are two significant challenges for creating robust image segmentation methods. Using sensors for gesture recognition has the advantage that movement information is provided directly.

Gesture recognition using sensors has been studied for some time. Several systems using complex arrays of devices have been deployed for real world use [6] and others restricting themselves to constrained environments for human computer interaction [7], [8].

In a previous approach [9], we considered the problem of gesture recognition from a continuous stream, but used a standard hidden Markov model and considered any region of movement in the stream as a candidate gesture. The approach was unable to deal with unknown movements and was sensitive to the threshold for gesture duration. In the following, we describe how these problems were overcome.

3 Hierarchical Modeling of Gestures

Gestures can be considered to exist at multiple levels in a hierarchy, where simple movements are grouped into more complex movements and complex movements are grouped into ordered sequences. The advantage of hierarchical modeling is this temporal decomposition of gestures. The classification stage becomes more manageable for increasingly complex gestures as the dynamics of the gesture are explicitly encoded. Not only does grouping allow for segmenting a gesture into its subparts, hierarchical modeling allows new gestures to be learnt on-line by reusing subparts from already known gestures.

In previous work [10], we proposed an extension to the hidden Markov model for recognising hierarchical gestures. Our extension was applied to recognising Kung–Fu gestures where each individual move was considered a sub-gesture. The extension was capable of representing the hierarchy of gestures, however, higher levels in the model had to be hand crafted. Another recent extension to the hidden Markov model, the

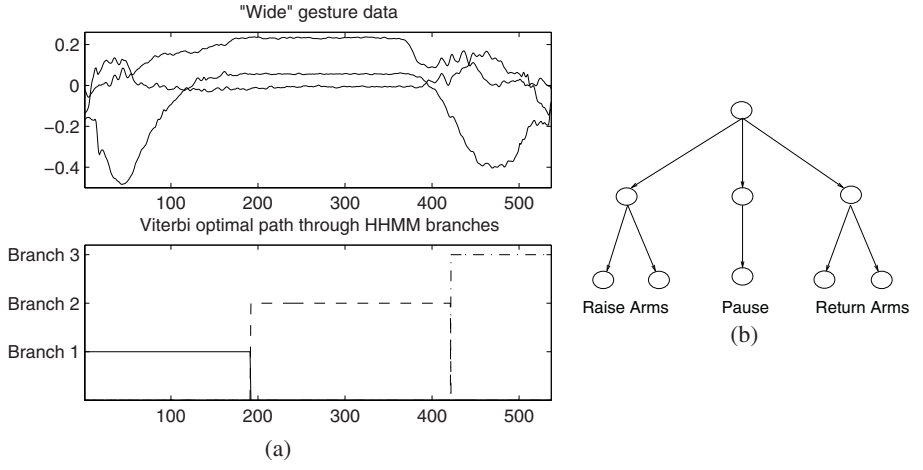

Fig. 1. "Wide" gesture data with corresponding viterbi path (a) and its HHMM representation (b).

HHMM, has been proposed [11]. The HHMM has a distinct advantage over our previous approach in that the model parameters can be estimated using a modified Expectation Maximisation algorithm. New gesture sequences can then be learnt on-line using the HHMM. For example, Figure 1 shows accelerometer data for the right arm and the corresponding optimal state path at the top level of the gesture HHMM. The movement corresponds to a "Wide" umpire gesture from the game of cricket whereby the umpire lifts both arms to the side until they are horizontal, holds that position for a short time so other officials can see the movement, then returns the arms back to the body. The HHMM allows us to explicitly model these three components of the gesture by considering the complete gesture as a sequence of three parts. The HHMM subtrees are trained independently using standard hidden Markov model techniques, then the parameters are merged into the HHMM structure. Instances of the complete gesture (combined subparts) are then used to train the higher levels of the HHMM. The Gaussian mixture components at the lowest level are not re-estimated.

Figure 1(b) shows the hierarchy used for the gesture, where the left-most branch is the raising of the arms, the centre branch is the pause, and the right-most branch is returning the arms to the body. The optimal (Viterbi) path through the top level in the HHMM illustrated in the lower portion of Figure 1(a) shows how the HHMM can automatically segment gestures at different levels. Time 1 through to 192 is recognised as the "raise arms" portion of the gesture, then from 193 to 422 is recognised as the "pause" portion of the gesture, then finally at time 423 to 537 is the "return arms" portion of the gesture.

4 Continuous Gesture Recognition

Any real application of gesture recognition has to be able to deal with segmenting gestures from a continuous stream, then classifying these candidate gestures to give an appropriate label. This process is similar to speech recognition systems whereby spoken

words are detected and recognised by the system. An approach to detecting unknown words in speech is applied in our gesture recognition system. A "filler" (or "null") model is created, being essentially an average of all other models in the system [12]. In our case, the model is simply trained on all available gesture classes.

Let $P(X|M_i)$ be the likelihood of the observation sequence given the gesture models M_i. Let $P(X|F)$ be the likelihood of the observation sequence given the filler model F. Then the sequence X is classified as model M_i if

$$FR = \frac{P(X|M_i)}{P(X|F)} \geq A \quad \text{and} \qquad (1)$$
$$P(X|M_i) > P(X|M_j)(i \neq j)$$

where A is a suitably chosen threshold (1.2 in our case) and FR is the filler ratio. Equation 1 then provides a way of determining the significance of an observation sequence based on an appropriately chosen threshold.

4.1 Extraneous Movement

Sports officials are obviously going to perform movements which will not be modeled by the system. Bending down to tie a shoe lace, for example, gives very little information on what is going on in the game. These extra movements should therefore be identified and potentially ignored. It is quite possible that the likelihood of some unknown movement will approach the likelihood of known movements, thus there must be some kind of confidence on how well an observation matches the set of known gestures. Using the filler model allows us to compare different observations relative to each other and provides us with a measure of confidence on how well a model matches the observation sequence.

Potentially more than one model will exceed the threshold for the filler model, however in this work, we simply take the maximum. The difference between the most likely model and the second most likely model is not taken into account.

4.2 Segmentation

Following our previous approach [9], the gestures are first segmented by using the magnitude of accelerometer data (see Section 5 for details). A Gaussian distribution modeling the magnitude of gravity is used to detect periods of movement over a sliding window. The likelihood of the Gaussian is used to make a binary decision on whether the window contains movement. In some cases, when there is little acceleration, spurious responses to the Gaussian distribution can occur. For example, the sequence 1,1,0,1 (where 1 is movement and 0 is no movement), may result, however this is not consistent labeling since there is a large degree of overlap $(48/144)$. Thus this sequence is replaced by the sequence 1,1,1,1. Similarly, if the sequence 0,0,1,0 occurs, it is replaced with the sequence 0,0,0,0. These two filters are ensuring that contiguous regions of data have consistent labeling over a sliding window. Figure 2 shows an example sequence of gestures with the corresponding movement decision below.

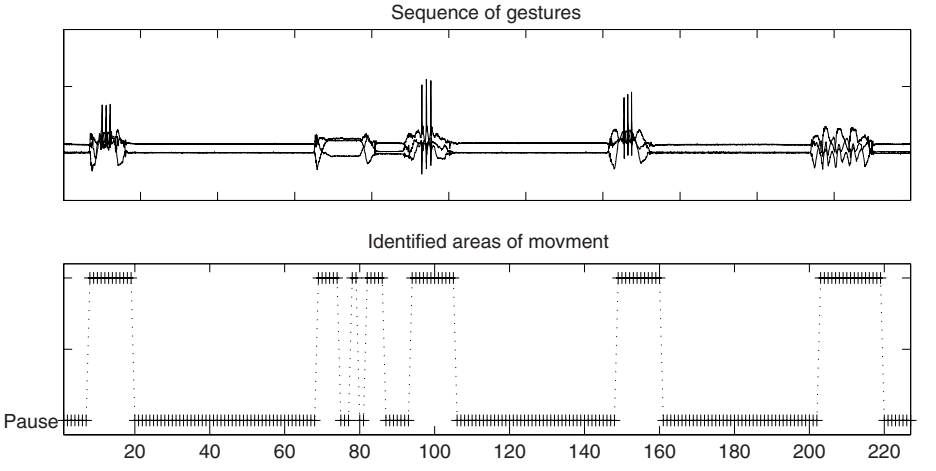

Fig. 2. Segmentation of gesture data using Gaussian of stationarity.

Once the regions of movement indicating candidate gestures are detected, the regions must be classified. Since many sport umpire gestures contain pauses such as the cricket example from Figure 1, there needs to be a method for grouping adjacent regions of identified movement including the pause periods. The problem is that a pause can be either a valid sub-gesture (as part of a gesture) or a pause between gestures. Our approach to overcoming this is by using a conservative estimate for the maximum length of a gesture (10sec) and grouping all detected gestures and pauses in the window. This window is then iteratively reduced in length removing the last region of movement at each step. In Figure 2, for example, the first gesture starting after time 50 would have four candidate regions for grouping (times 69 to 106, 69 to 86, 69 to 79, and 69 to 75). The algorithm proceeds as follows: for each period of movement ahead in time (up to 10sec) of the start of a candidate gesture, calculate the likelihood of each model for that region. After all candidate regions have been identified and their corresponding model likelihoods calculated, they are normalised by the filler ratio. The region and model corresponding to the maximum of the calculated ratios is considered the gesture for that region. Using a filler model to compare different length observations for accurately finding segmentation endpoints is novel and has worked well on our data and example domain. Without a filler model, different length observations can not be compared across models as the HMM likelihood function is non-linear. The segmentation example referred to in Figure 2 is able to be correctly segmented with our approach.

5 Experimental Results

The architecture of the system is illustrated in Figure 3. As video data is recorded, the movement of key actors wearing the accelerometer wrist bands is also recorded. The accelerometer data is analysed then segmented and classified using the segmentation and classification algorithms described. If a gesture of interest was detected, the video

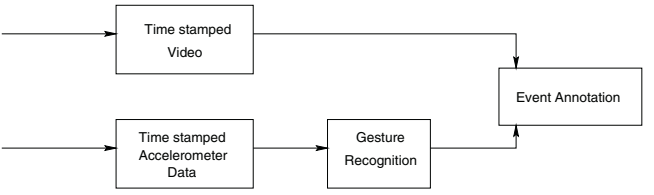

Fig. 3. System Overview.

Table 1. Results for the four controlled sequences.

	Known	Unknown	Recall
Known	40	0	100%
Unknown	2	7	77.8%
Precision	95%	100%	

at the time the gesture was performed is then annotated with the event indicated by the gesture.

Our accelerometers can measure acceleration in 3–D space. They are housed in a small wrist watch sized enclosure worn in the form of a wrist band. Obviously the recognition performance of the system could suffer if the band was worn in grossly different orientations on the wrist, thus we treat the band like a watch, where the face of the enclosure is in a similar direction each time the band is worn. The implementation measures acceleration of up to $\pm 2g$ at 150 samples/second. The feature set used is standard deviation and root mean square for each of the three directions of acceleration. Features are calculated using a window size of 48 samples ($\approx$ 300ms) with a 12 sample overlap. The window size for segmenting regions of movement is 144 samples with a 48 sample overlap.

The cricket umpire gestures we recognise are: *Dead Ball* – sway both arms in front of the body, *Four* – wave the right hand across the body, *Last Hour* – point to the watch on the raised left arm and tap it, *Leg-Bye* – tap the raised right knee, *No Ball* – extend right arm to the side, *One Short* – tap right shoulder with right arm, *Out* – raise arm in front of body with index finger extended, *Penalty Runs* – grasp left shoulder with right hand, *TV Replay* – Outline a rectangle with both hands, and *Wide* – extend both arms out to the sides of the body.

5.1 Controlled Sequences

To initially test our technique, four sequences containing each gesture and a number of unknown movements are recorded. In these sequences, the actor intentionally performs a gesture, either known or unknown, then pauses for some random period, then performs another gesture. In each of the four sequences, each of the 10 known gestures are performed. For the first sequence, one unknown gesture is performed, for the second sequence, two unknown gestures are performed, then for the third and fourth sequences, three unknown gestures are performed. The unknown gestures include the

Table 2. Results for three match segments.

	Known	Unknown	Recall
Known	12	0	100%
Unknown	16	51	76%
Precision	42%	100%	

actor scratching their head, picking up pens from a desk, and typing for a short duration on a computer keyboard. Some of the movements were intentionally performed with very little time between them to test the performance of the segmentation algorithm. Figure 2 shows a portion of one such sequence. Table 1 shows the classification results for the four sequences. The algorithm was able to correctly identify the start and end regions of all gestures in all sequences and classify all known gestures correctly. Two unknown movements were incorrectly classified as known movements.

5.2 Cricket

Two 15 minute portions of an Australia versus England test match and one 15 minute portion of an England versus Pakistan test match were recorded. In all three sequences, the actor mimicked the umpire and performed both known and unknown movements to better represent real world data. The unknown movements include actions such as walking, bending over to pick up something, and passing objects to players. The results are again shown as a confusion matrix in Table 2. When the gesture is known, the system performs well, classifying all gestures correctly. The table shows, however, that unknown gestures are frequently detected as known gestures. A large proportion of these unknown gestures being incorrectly classified as known gestures can be partially explained since the umpire consistently does an action which resembles one of the known gestures.

Typically a cricket umpire stands behind the wickets with their hands behind their back. When the ball is "in play", it is common for the umpire to take their hands from behind their back and start to walk away from the wickets to prevent interfering with the game play. The "Dead ball" gesture starts in a very similar manner and is frequently falsely detected as occurring in the sequences. Currently we are looking at reducing this by using more meaningful criteria than a simple threshold.

5.3 Video Labeling

Table 3 shows the indexing performance for one of the mimicked segments. The table lists the start (t_{1g}) and end (t_{2g}) times of the known gestures in the ground truth and the start (t_{1d}) and end (t_{2d}) times detected by the system. The difference in start time (δt_1) and gesture length ($\delta(t_2 - t_1)$) is also listed. Both the video frame rate and the size of the overlap in the sliding window for movement detection affect the accuracy of the system. The video is recorded at 25fps and the size of the appended data to the sliding window is 48 samples (0.32sec). The table shows that the first three gestures are detected consistently around 1.2 seconds before the gesture occurred. The final gesture

Table 3. Gesture Results for a match segment.

Gesture	t_{1g}	t_{2g}	t_{1d}	t_{2d}	δt_1	$\delta(t_2 - t_1)$
Wide	90.8	95.6	89.6	94.22	1.2	0.04
Four	96.8	102.56	95.68	100.80	1.2	0.64
Four	319.4	323.36	318.08	323.52	1.3	0.52
Wide	722.8	727.76	717.44	726.08	5.3	3.68

however was detected 5.3 seconds early. The reason the start time of the final gesture was incorrect is that an unknown movement was performed just before the gesture and the segmentation algorithm grouped the adjacent regions together considering them as one complete gesture. This failure of the segmentation algorithm was the only time it incorrectly identified the regions corresponding to a complete gesture for all tested sequences.

The results show that the system performs well overall with the exception of handling unknown movements which have similarities to known movements. Detecting movement is robust, however utilizing the filler ratio requires further investigation for deciding when a known gesture occurs.

6 Conclusions

We have presented results on our approach toward automatic sports video annotation in which we solve the problem of automatic segmentation and robust gesture classification using a hierarchical hidden Markov model in conjunction with a filler model. The hierarchical model allows us to consider gestures at different levels of abstraction and the filler model allows us to handle extraneous umpire movements. We have used a variable sized window to group regions of movement to overcome the problem of recognising gestures which contain pauses as part of the gesture. Pauses between regions of movement need to be regarded (in some cases) as part of a gesture and in other cases, as gaps between gestures. The concept of a filler model from speech recognition is used in our system to aid in detecting unknown movements and classifying them accordingly. Further work will be in investigating the filler model ratio to provide more insight than the current approach which is simply a threshold.

References

1. Gong, Y., Sin, L.T., Chuan, C.: Automatic parsing of TV soccer programs. In: Proceedings of the IEEE International Conference on Multimedia Computing and Systems. (1995) 167–174
2. Zhou, W., Vellaikal, A., Kuo, C.: Rule–based video classification system for basketball video indexing. In: ACM Multimedia 2000, Los Angeles, USA (2000) 213–216
3. Pan, H., Beek, P.V., Sezan, M.I.: Detection of slow-motion replay segments in sports video for highlights generation. In: Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing. (2001) 1649–1652
4. Yang, M.H., Ahuja, N.: Extraction and Classification of Visual Motion Patterns for Hand Gesture Recognition. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Santa Barbara, California (1998) 892–897

5. Becker, D.A.: Sensie: A real–time recognition, feedback and training system for T'ai Chi gestures. Technical Report 426, M.I.T. Media Lab Perceptual Computing Group (1997)
6. VPL DataGlove: VPL DataGlove. VPL Research (2000)
7. Benbasat, A., Paradiso, J.: An inertial measurement framework for gesture recognition and applications. In Wachsmuth, I., Sowa, T., eds.: Gesture Workshop. Springer-Verlag (2002) 9–20
8. Hinckley, K., Pierce, J., Sinclair, M., Horvitz, E.: Sensing techniques for mobile interaction. Symposium on User Interface Software and Technology, CHI Letters **2** (2000) 91–100
9. Chambers, G.S., Venkatesh, S., West, G.A.W., Bui, H.H.: Segmentation of intentional human gestures for sports video annotation. In: International Conference on Multimedia Modelling. (2004) 124–129
10. Chambers, G.S., Venkatesh, S., West, G.A.W., Bui, H.H.: Hierarchical recognition of intentional human gestures for sports video annotation. In: International Conference on Pattern Recognition. (2002) 1082–1085
11. Fine, S., Singer, Y., Tishby, N.: The hierarchical hidden markov model: Analysis and applications. Machine Learning **32** (1998) 41–62
12. Williams, G., Renals, S.: Confidence measures for hubrid hmm/ann speech recognition. In: Proceedings of Eurospeech 1997, Rhodes (1997) 1955–1958