Human Action Recognition with Primitive-based Coupled-HMM

Haibing Ren, Guangyou Xu

Department of Computer Science and Technology, Tsinghua University, State Key Laboratory of Intelligent Technology and Systems Beijing 100084, PR China renhb98@mails.tsinghua.edu.cn

Abstract

This paper presents a new approach named Primitive-based Coupled-HMM for human natural complex action recognition. First, the system proposes a hybrid human model and employs 2-order B-spline function to detect the two shoulder joints in the silhouette image to obtain the basic motion features including the elbow angles, motion parameters of the face and two hands. Then, Primitive-based Coupled Hidden Markov Model (PCHMM) is presented for natural context-dependent action recognition. Last, comparison experiments show PCHMM is better than the conventional HMM and coupled HMM.

1. Introduction

The Smart classroom is a project of intelligent environments for tele-education. This paper focuses on the teacher’s upper-limb action recognition to understand the teacher’s intention for intelligent cameraman and virtual mouse system. The framework of the recognition system is as following:

![Figure 1. System framework.](image)

In the smart classroom, there are a lot of cameras around the teacher. By the face detection module [3], the two neighboring cameras that face the teacher are auto-selected. Therefore, only the frontal actions are considered in this paper. By face recognition module [6], the teacher’s identity is recognized and his personal information is retrieved in the database, which is very important to get more 3D motion feature in section 2 and adapt to his habit in further application.

1.1 Related work

In most previous research on action recognition, the actions are confined to a predefined command set, which requires that the subjects are well trained and the motions are uniform. The features and recognition algorithms are totally data-driven without any information of high-level feature or context information.

For example, in [7], moment-based features are extracted from multiple views of motion energy images (MEI) and motion history images (MHI), and template matching algorithm are employed to recognize the aerobics exercises and the well performed moves in the KidsRoom. And in [2], with the coordinates of the two hands, Hidden Markov Models are used to recognize American sign languages. In [9], with 3D trajectories of the two hands, coupled Hidden Markov Models are presented to recognize 3 kinds of T’ai Chi Ch’uan.

However, the teacher’s natural action recognition is much more difficult than the above. The motion is more natural, complex and dependent on the context and scenario. No subject would be trained and everyone has his own habit.

1.2 Our approach

In psychological research on action recognition, it was found that motion models in subjects mind are not the motion parameters such as the parameters of position and motion speed but merely some characteristic features, like the relative position of hands and face, the relation between the moving hand and the scenario, etc.

Based on this principle, this paper presents a new system to recognize the teacher’s natural complex action in the smart classroom. And much of it has general sense and can be used in other areas.

First, this paper gives the hybrid human model to obtain the basic motion features including motion feature of the face, two hands and the two elbow angles. Then, a recognition algorithm, named Primitive-based Coupled Hidden Markov Model (PCHMM), is presented to recognize the subject’s natural complex action.

As a totally data-driven algorithm, the conventional CHMM is not suitable for natural complex action recognition because the feature dimension is too large and the within-class scatter is too much. Unfortunately, it is
impossible to get enough training samples containing everyone’s every type of action. Unlike conventional CHMM, PCHMM is an approach with high-level features and context information, which need less training samples and could diminish the within-class scatter greatly. This paper is arranged as follows: section two describes the hybrid human model and the basic motion feature estimation; and section three gives the PCHMM for upper-limb action recognition; section four is the experimental results and last is the conclusion.

2. Hybrid human model

In many action recognition systems, only the trajectories of two hands (or only 1 hand) are extracted as motion features, hence much ambiguity in recognition. In contrast, this paper presents the hybrid human model as Figure 2 to get more explicit 3D motion information. This model includes the face, the trunk, two arms and two hands, which are indispensable parts for any person and can be easily detected, and with which more information can be given to reduce the complexity and improve the robustness greatly.

2.1 Context

In this paper, the context is a generic term, including the teacher’s personal information indexed by the face recognition result. Though only a little is useful in this paper, the teacher’s personal information is very important for further applications such as human 3D modeling, more precise motion estimation, habit self-adaptation, etc.

The context also includes some scenario information, such as the position of some objects on the desk. And with the result of previous action understanding (taking objects from the desk, laying down objects, etc), the variant $b_{object}$ in the context represents the object in the subject’s hand.

2.2 Shoulder joints detection

With the teacher’s silhouette image obtained by background subtraction, this paper gives an effective approach to detect the two shoulder joints. First, the trunk area is segmented in the silhouette image as following:

$$\text{Area}_{\text{Trunk}} = C(\text{Area}_{\text{Silhouette}} - \text{Area}_{\text{Face}}) - \text{Area}_{\text{arm}}$$  \hspace{1cm} (1)

Where the $\text{Area}_{\text{Silhouette}}$ is the silhouette area as Figure 3(a), $\text{Area}_{\text{Face}}$ is the face area, $\text{Area}_{\text{arm}}$ is the two arm areas (including two hands) and $C$ is an operator to get the maximum connective area. The histogram of $C(\text{Area}_{\text{Silhouette}} - \text{Area}_{\text{Face}})$ on the $X$ coordinate as Figure 3(c) can be looked as 5 lines as Figure 3(d). The areas covered by $l_1$ and $l_2$ are considered as $\text{Area}_{\text{arm}}$. By deleting the two arm areas, the left is $\text{Area}_{\text{Trunk}}$ as Figure 3(e). The parameters of $l_i$ ($i = 1,...,5$) is estimated to satisfy the following constraint:

$$\min \left[ \int_{s=0}^{1} \left| (H_x(s) - F_x(s)) \right| + \int_{s=0}^{1} \left| (H_y(s) - F_y(s)) \right| \right]$$ \hspace{1cm} (2)

Where $s$ is arc length normalized to $[0, 1]$ , $H(s) = (H_x(s), H_y(s))$ is the histogram function, $F(s) = (F_x(s), F_y(s))$ is the function composed of the 5 lines. The equation (2) is a global optimization problem with hyper parameters. In this paper, 1-order B-spline with 5 segments is employed to get the solution and the result is as the Figure 3(d), which is very precise, robust and only costs about 2.56ms/histogram.

Last, two special corner operators are used to detected the two shoulder joints in the ROI (region of interest) which is determined by the position and the size of the $\text{Area}_{\text{Face}}$. With the feature extracted around the shoulder joints, the corresponding one in the other image is located by matching algorithm.

When an arm is in front of the trunk, there may be some ambiguity between $l_1$ and $l_2$ (or $l_4, l_5$). But in this case, $l_1$ and $l_2$ are very small in the histogram and no matter how much the error of $\text{Area}_{\text{arm}}$ is, the deviation of $\text{Area}_{\text{Trunk}}$ will be very little. As the detection of the shoulder joints is based on the global information, the results are very robust.
2.3 Basic motion features

With the position of the face, two hands and two shoulder joints in stereo images, their 3D coordinates can be obtained. Then elbow joint \( \theta \) \((0 \leq \theta \leq 90^\circ)\) are calculated with the integration of shoulder joint 3D coordinate and the length of the upper, lower arms. In this paper, the elbow angle is more important than elbow joint 3D coordinate. The reason is:

1. Because human 3D modeling is time-consuming and unstable, it is nearly impossible to estimate the elbow joint 3D coordinate precisely and robustly without any markers.
2. For action recognition, it is unnecessary to get the elbow joint 3D coordinates, because the position of elbow joint is meaningless in most human motion.
3. The elbow angle is significant to represent the arm state.

Therefore, the basic motion feature for each hand is obtained as following:

\[
(P_{hand}, V_{hand}, A_{hand}, P_{face}, V_{face}, \theta_{elbow}, b_{object})
\]

where \(P_{hand}, V_{hand}, A_{hand}\) is the hand 3D position, 3D velocity and 3D acceleration respectively. And \(P_{face}, V_{face}\) is Face 3D position and 3D velocity, \(\theta_{elbow}\) is the elbow angle and \(b_{object}\) is a variant in context. From equation (2), the following could be got:

\[
V_{hand} = \partial P_{hand} \quad A_{hand} = \partial V_{hand} \quad (3)
\]

\[
V_{face} = \partial P_{face} \quad (4)
\]

Due to different teachers and different actions, there is much within-class scatter in the 17-dimension basic motion features. And it could not be used directly for action recognition.

3. Primitive-based Coupled Hidden Markov Model

This paper introduces the primitive features to the conventional coupled-HMM and calls it Primitive-based Coupled Hidden Markov Model.

3.1 Primitive features

The states in vision-based action recognition often have definite meaning and much clear segmentation. It is because each state has some unambiguous features, called primitive features or primitives in this paper. Each primitive feature \( \lambda \) is represented by a Gaussian Mixture Model (GMM) and the distribution density is as the following:

\[
P(p \mid \lambda) = \sum_{i=1}^{G_n} G_i(p)P(i) \quad (5)
\]

where \( p \) is a primitive variant, \( G_i(p)(i=1,\ldots, G_n) \) is a Gaussian model with mean \( \xi_i \) and covariance matrix \( \sigma_i \), which are hyperparameters of the distribution, \( G_n \) is the Gaussian model number, \( p(i) \) is the weight function for each Gaussian model. The parameters of \( G_i(p) \) and \( p(i) \) could be estimated by Expectation Maximum (EM) algorithm.

3.2 Representation of the states

The states in this paper are strictly defined by some primitive features and corresponding weight. These primitive features of state \( S \) are supposed to be independent of each other. The observation densities function of \( S \) is as following:

\[
P(O \mid S) = P(P \mid S) = \sum_{i=1}^{P_n} P(p_i \mid \lambda_i)w_i \quad (6)
\]

where \( O \) is the observation(basic motion feature in this paper), \( P_n \) is the primitive feature number of \( S \), \( \lambda_i \) is the \( i \)th primitive feature \((i=1,\ldots, P_n)\), \( w_i \) is the corresponding weight for \( \lambda_i \), \( P = \{p_1,\ldots,p_{P_n}\} \) is the primitive set of \( S \). And the relation of \( p_i \) and \( O \) can be described as the equation (7):

\[
p_i = f_i(O,S,\text{Context}) \quad (7)
\]

where \( \text{Context} \) is the context information and \( f_i \) is the function to extract the \( p_i \) from \( O \), \( S \) and \( \text{Context} \). In
this paper, the weight vector \( W = (w_1, w_2, \ldots, w_N) \) is estimated by the following:

\[
\arg \max_w \left\{ \sum_{i=1}^{N} P(O_i | S) \right\}
\]

\( w(i) = 1 \) 

where \( w_i, S, P_n \) are the same as equation(6), \( O_i \) is the \( i \)th training sample for state \( S \) and \( T_n \) is the number of the training samples. Here, the weight vector \( W \) is estimated by maximum likelihood [4].

### 3.3 Framework of PCHMM

For each hand, the basic motion feature from time 1 to \( T \) is considered to be 1-order Markov chain. And suppose that the relation between the two hands satisfies PCHMM, as the Figure 4.

![Figure 4. PCHMM structure.](image)

where the superscript \( t, t+1, t+2 \) in the structure means at time \( t, t+1, t+2 \) respectively, \( L' \) is the left hand state at time \( t \), \( A' \) is the basic motion feature of the left hand at time \( t \), \( P' \) is the primitive feature of left hand motion, \( W' \) is the weight vector for \( L', R' \) is the right hand state at time \( t \), \( B' \) is the basic motion feature sequence of the right hand at time \( t \), \( P'' \) is the primitive feature of the right hand motion, \( W'' \) is the weight vector for \( R' \).

\( L'_t = \{L', t = 1, \ldots, T\} \) is the left hand state sequence, \( A'_t = \{A', t = 1, \ldots, T\} \) is the basic motion feature sequence of the left hand. \( R'_t = \{R', t = 1, \ldots, T\} \) is the right hand state sequence, \( B'_t = \{B', t = 1, \ldots, T\} \) is the basic motion feature sequence of the right hand. The likelihood function of PCHMM with the basic motion feature \( A'_t \) and \( B'_t \) is:

\[
P(A'_t, B'_t | \Theta) = P(L'_1) * P(R'_1) * P(A_t | L')
\]

\[
= \prod_{i=2}^{T} \left[ P(A_i | L_{i-1}, R_{i-1}) * P(R_i | L_{i-1}, R_{i-1}) \right]
\]

\[
P(A'_t, B'_t | \Theta) = P(A_t | L') * P(B_t | R')
\]

Where the PCHMM parameter set \( \Theta \) contains the prior probabilities \( P(L'_1) \) and \( P(R'_1) \) for the two Markov chains, the observation densities function \( P(A_{i-1} | L_{i-1}) \) and \( P(B_{i-1} | R_{i-1}) \), the transition probabilities \( P(L'_{i-1}, R'_{i-1}) \) and \( P(R'_{i-1}, R'_{i-1}) \). The prior probabilities \( P(L'_1) \) and \( P(R'_1) \) are supposed to be equal for each model. With forward-backward Viterbi algorithm, the parameters set \( \Theta \) of the model can be estimated as following:

\[
\arg \max_{\Theta} \sum_{i=1}^{S} P(A_{i1}^T, B_{i1}^T | \Theta)
\]

where \( s_n \) is the sequence number for training, \( (A_{i1}^T, B_{i1}^T) \) is the \( i \)th training sequence.

### 4. Experimental results

For the teacher in the smart classroom, there are totally 7 kinds of natural actions to be recognized:
- Taking objects from the desk
- Returning objects to their former places
- Pointing to the students
- Pointing to the blackboard (virtual mouse)
- Communicating with the students
- Explaining objects
- Drinking water

For each action, there are 50 samples. Comparison experiments are done among HMM, conventional CHMM and PCHMM and the result is as following:

![Figure 5. The comparison among HMM, conventional CHMM and PCHMM.](image)

where the x coordinate is the size of the training data set and all the testing sets are the whole data set. It shows PCHMM is the best among the three algorithms, especially with less training data.
Another comparison experiments on the elbow angles are carried out. And the result is as following:

![Figure 6. The comparison on the elbow angles.]

where ‘With’ means motion features include the elbow angles and ‘Without’ means the reverse. The Figure 6 shows the basic motion features with the elbow angles perform much better than the one without them. Though the elbow angles are not very precise, they are very important features to represent the arm state. At the same time, it shows the face and two hands motion features aren’t sufficient for complex action recognition.

5. Conclusion

This paper presents a framework for the teacher’s complex action recognition in the smart classroom. With the Hybrid Human Model, basic motion features are extracted which include the two elbow angles and the motion features of the head and two hands. Primitive-based Coupled-HMM are used for recognition. And the encouraging experiment result shows the PCHMM is very robust and can obtain better result especially in the case of less training samples.

References