A Fuzzy Qualitative Approach to Human Motion Recognition

Chee Seng Chan, Honghai Liu, David Brown and Naoyuki Kubota

***Abstract*— The understanding of human motions captured in image sequences pose two main difficulties which are often regarded as computationally ill-defined: 1) modelling the uncertainty in the training data, and 2) constructing a generic activity representation that can describe simple actions as well as complicated tasks that are performed by different humans. In this paper, these problems are addressed from a direction which utilises the concept of fuzzy qualitative robot kinematics [9]. First of all, the training data representing a typical activity is acquired by tracking the human anatomical landmarks in an image sequences. Then, the uncertainty arise when the limitations of the tracking algorithm are handled by transforming the continuous training data into a set of discrete symbolic representations - *qualitative states* in a quantisation process. Finally, in order to construct a template that is regarded as a combination ordered sequence of all body segments movements, robot kinematics, a well-defined solution to describe the resulting motion of rigid bodies that form the robot, has been employed. We defined these activity templates as qualitative normalised templates, a manifold trajectory of unique state transition patterns in the quantity space. Experimental results and a comparison with the hidden Markov models have demonstrated that the proposed method is very encouraging and shown a better successful recognition rate on the two available motion databases.**

I. Introduction

Recognising human activity from a stream of video sequences is important for a number of applications, in particular surveillance, video understanding and human-computer interaction. To this extent, significant amounts of research have been conducted to represent, annotate and recognise the activities performed by humans. Presently, probabilistic graph models including both temporal sequential models such as Hidden Markov Models (HMMs) and static causal models such as Bayesian belief networks have received enormous attention from various communities for modelling and recognising human activities. A recent review [4] even reported the field almost entirely in terms of these methods.

In probabilistic graph models, given the training data extracted from video, either as motion trajectories or labelled discrete events, activities are modelled as a set of structured states in a state space. These states are linked by a set of causal or temporal connections referred to as the structure of the model. The model requires both the determination of the states, through clustering of training data sets, and the discovery of the underlying structure performed by the factorisation of the state space.

For instance, Yamoto et al. [22] made use of HMMs to recognise human actions based on low resolution image intensity patterns in each frame. These patterns were passed to a vector quantiser, and the resulting symbolic sequence was recognised using the HMMs. Bobick [2] and Campbell [5] mapped Cartesian tracking data (captured from sensors on body joints) onto a body hierarchy for activity recognition. The trajectory data of the joints are represented in a high dimensional phase space, and points in this space (or one of its subspaces) are employed to recognise activity. The basic idea in both approaches was to use the dynamic characteristics of motion to achieve activity recognition. Wilson and Bobick [21] proposed a modified HMMs approach to activity recognition. They introduced the term parametric gestures defined as gestures that exhibit a systematic spatial variation. As an example, the paper cited a pointing gesture, where the relevant parameter is a two-dimensional direction. The standard HMMs method of gesture recognition was extended by including a global parameter-driven variation in the output probabilities of the HMMs states. Then they formulated an Expectation-Maximisation (EM) method for training this parameter driven HMMs. During testing, a similar EM algorithm simultaneously maximises the output likelihood of the parametrically driven HMMs for the given motion sequence, whilst estimating the quantifying parameters. EM training is based on learning from samples, and the parametric aspect in terms of direction is based on prior knowledge of certain gestures. This approach assumes the availability of pre-segmented gestures. Oliver et al. [13], on the other hand, employed coupled hidden Markov model [3] to model pedestrian activity for surveillance and analysing actions which occur between two pedestrians. In this model, two (or more) HMMs are coupled, with the state of each

Fig. 1. Sample activities used from the database provided by [1], [16]. Trajectories from seven landmarks (left/right shoulders, left/right elbows, left/right knees and hip) on the human body are quantised and input to our method.

Chee Seng Chan, Honghai Liu and David Brown are with the Institute of Industrial Research, University of Portsmouth, Portsmouth PO1 3QL, U.K. (e-mails: {cheeseng.chan;honghai.liu;david.j.brown}@port.ac.uk.), Naoyuki Kubota is with the Department of System Design, Tokyo Metropolitan University, Tokyo 192-0397, Japan. (e-mail: kubota@comp.metro-u.ac.jp)

978-1-4244-1819-0/08/$25.00©2008 IEEE

at time *t* affecting the state at time *t+1*. They trained their model on synthetic data, and a mixture of synthetic and real data.

While all these approaches have demonstrated success in modelling and recognising complex activities, there is a tendency to use the parameterisation as a 'black box'. That is, these approaches depend on the probabilities and extensive training to recognise all of the activities. Therefore, one needs to have a large number of training sequences for each activity to be successfully recognised. Bear in mind that the complexity of Bayesian networks are proportional to its number of states. Moreover, we also notice that most researchers had to rely on cumbersome tracking algorithm to obtain the training data. For example, a number of systems are based on incremental updates or the searching around a predicted value [7], [19], [20]. Many of these tracking systems will fail due to occlusion, bad predictions or a change in the frame rate and the risk of accumulating errors due to the incremental procedure. Nevertheless, a standard particle filter has a computational complexity of *O(2N)* and time complexity of $N\sum_{k=1}^{m}\tau_k$ where *N* is the number of particles and τ_k is the cost of calculating $p(z_k|x)$. Whereas methods that do not use such approaches usually rely on the accuracy of motion sensors, but seldom provide a measure of confidence of the results which is crucial for discriminating similar events in a noisy environment.

Therefore, it is desirable to model the uncertainty in the training data, and yet still to be able to describe a simple action as well as complicated activities performed by humans. In this paper, we approach this by utilising the fundamental concepts of fuzzy qualitative robot kinematics [9] as our methodology. The basic idea is firstly that we consider the human body as being composed of physical parts connected by joints, for instance, the upper arm is connected by the elbow joint to the lower arm, and the lower arm is connected to the hand by the wrist joint. Each part can move independently, and hence can exhibit an independent degree of activity. From here, a set of spatio-temporal training data from the different human body parts in a video stream are extracted by a simple tracking algorithm. In order to model the uncertainty that arise due to the limitation of the tracking algorithm, these continuous training data are then transformed into a set of discrete symbol representations - *qualitative states* thru a quantisation process. Each of these states are derived and normalised in a modified unit circle [10]. Depending upon the corresponding region in which the quantitative dynamic characteristic of motion data resides, the data is assigned to a particular state.

An activity is defined as a combination of ordered sequences of all body movements or states, restricted by the physical anatomical limits [2], [8]. In this paper, our second objective is to build an activity representation that is based on this taxonomy. The concept of fuzzy qualitative robot kinematics [9] which is a well-establish solution in robotics industry to describe the motion between the joints of the manipulator and resulting motion of the rigid bodies which form the robot is therefore, exploited. We defined these templates as Qualitative Normalised Templates (QNTs), a manifold trajectory of unique state transition patterns in the quantity space.

Empirical results on the two available databases and a comparison with the HMMs approach have shown that our proposed method 1) can cope with uncertainty in the training data thus ruling out the need for cumbersome tracking algorithms and 2) utilisation of the robot kinematics solution which is a well establish approach in robotic industry to design the path planning of robotic manipulator to cluster a collection of motion from all the physical segments of human body to construct the QNTs, is significant over black-box methods such as neural network [11], HMMs [2], [3], [13], [21], [22] and etc.

The rest of the paper is structured as follows. Section II derives the qualitative normalised templates, in particular how both the quantisation process and fuzzy qualitative robot kinematics methodology are utilised to alleviate the problems. Section III presents the experimental results and a comparison with the HMMs. Section IV concludes the paper with discussions and future work.

II. Qualitative Normalised Templates

QNTs in contrary to probabilistic approaches [2], [3], [13], [21], [22] are a novel parametric activity representation that exploit the fundamental concepts of fuzzy qualitative robot kinematics [9] as a foundation. In this paper, an event Υ is defined as a temporal movement of body segment in a short time period and is represented by a state in the fuzzy unit circle. Whereas an activity **A** is defined as a combined ordered sequence of all the body segment movements over time, restricted by the motion constraints of the body. This is represented by the projection of state transition pattern in the quantity space. For example, the sequence of events (state changes) for a human walking include segment events for the foot, lower leg, and thigh, and joint events for the ankle, knee and hip. These sequences occur in the leg that is moving forward, while the leg that supports the body will show no events. A walking activity is defined as a combination of this sequence of events.

A. Acquisition of Training Data

To construct activity patterns, the training data that representing the activities should be acquired. In this paper, we considered the human body as a set of 13 anatomical landmarks, $i = \{1,2,\cdots,13\}$ as illustrated in Fig. 2. We choose these representations as they provide sufficient information about most of the activities. Moreover, modelling the human body as rigid parts linked in a kinematics structure is relatively easy to automatically detect and track in real videos, as opposed to the inner body joints which are more difficult to track.

For each activity performed by a human in a video stream, a tracking scheme [7] together with the proposed human model is performed to track the human. Formally, an activity **A** is a function of time. For each time T_c, we obtained

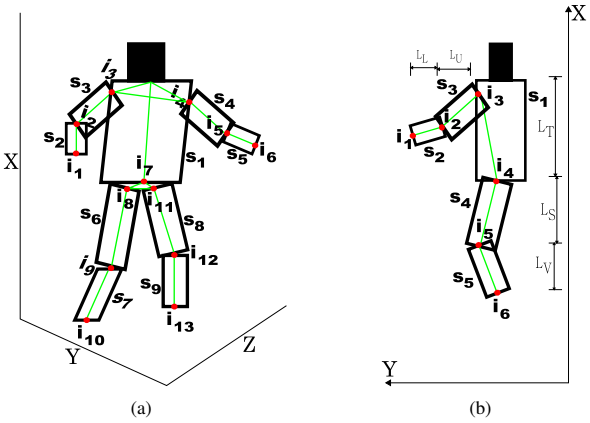


Fig. 2. The proposed human model. Each segment (limbs) of a human is represented by planar patches, connected by joints. (a) Full human body (b) Planar human body.

the trajectory corresponding to a point *i* that represents the human anatomical landmark as a sequence of locations: $[X^i_{T_c}\ Y^i_{T_c}\ Z^i_{T_c}\ T_c]^T$. These resultants have enabled us to measure the length of each human body part *L* and its respective joint angle θ. Therefore, an activity can be represented as: $\mathbf{A} = [\Upsilon_1, \Upsilon_2, \cdots, \Upsilon_j]$ where $\Upsilon_j = [L^i_{T_c}\ \theta^i_{T_c}]^T$ and $j \in i$. The construction of **A** must adhere to the ordering of *i*.

Note that, we are not attempting to solve the tracking problems in this paper. Thus, the details of the tracking algorithm will not be discussed, however we do refer readers to [7] for a detailed explanation. Also, any existing tracking algorithms can be employed as the front-end of the proposed method to obtain the training data. Fig. 1 shows the examples of sets of trajectories for different activities in the case of real videos.

B. Modelling Uncertainty in Training Data

Generally, there is a trade-off between tracking accuracies, computational complexity and time complexity. As described earlier, a standard particle filter has a computational complexity of *O(2N)* and time complexity of $N\sum_{k=1}^{m}\tau_k$. Whereas methods that do not use such approaches usually rely on the accuracy of motion sensors, but seldom provide a measure of confidence of the results which is crucial for discriminating similar events in a noisy environment.

In this paper, we proposed to model the uncertainty which arise due to the limitation of the computer vision tracking algorithm by exploiting the concept of fuzzy qualitative trigonometry [10]. Fuzzy qualitative trigonometry is a novel representation proposed by Liu and Coghill [10]. In the approach, axes in the conventional unit circle are replaced by unit quantity space; that is the Cartesian translation and orientation in the unit circle is replaced by a fuzzy membership function. The position state of a fuzzy qualitative point is defined by the projections of the point into fuzzy qualitative axes in the fuzzy qualitative unit circle. Four tuple fuzzy numbers [a, b, α, β] and its arithmetic [17], [18] are employed to describe the characteristic of each state in the fuzzy qualitative unit circle.

First of all, we considered the quantity space for the orientation and translation in the fuzzy qualitative unit circle as 16 and 12, respectively. That is, the description of its fuzzy qualitative orientation and translation are given by quantity space whose elements are a fuzzy membership function of real numbers of polar and Cartesian coordinates. For instance in Fig. 3, it can be seen that the 16 fuzzy numbers (states) in the quantity space of the orientation divides a quantitative range, e.g., [0 2π] into 16 qualitative regions with fuzzy boundary. These fuzzy numbers are the qualitative description of the quantitative orientation within a corresponding angular region.

Then, for each of the spatio-temporal training data (*L* and θ in this case) that were extracted from the different anatomical landmarks *i* at time T_c, we transform them into a set of discrete symbol representations, *qualitative states* in the unit circle by a quantisation process:

$$\begin{array}{l}\lim\limits_{s\to s_o} C_t(s=12) = QS(qp_l) \\ \lim\limits_{r\to r_o} C_o(r=16) = QS(qp_\theta)\end{array} \tag{1}$$

where s is the number of states that reside in the x-y translation while r is the number of states that reside on the orientation in the fuzzy qualitative unit circle. As s $\to s_o$ and R $\to r_o$, the limits of $C_t(s)$ and $C_o(r)$ will approach to a set of s_o qualitative states for a translation component and a set of r_o qualitative states for an orientation component. From this depending upon the corresponding region in which the quantitative dynamic characteristic of the training data resides, the corresponding symbolic representation of an activity **A** can be represented as:

$$\mathbf{A} = [\Upsilon_1, \Upsilon_2, \cdots, \Upsilon_j] \tag{2}$$

where

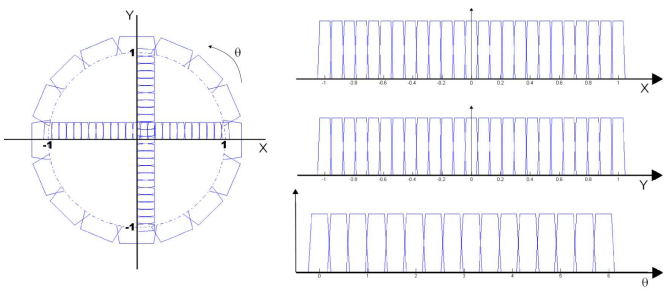
$$\Upsilon_j = [QS_{L^i_{T_c}}(qp_l)\ QS_{\theta^i_{T_c}}(qp_\theta)]^T \tag{3}$$

T_c is the time sequence of a video sequence, $QS_{L^i_{T_c}}(qp_l) \in$ {QS(1),QS(2),$\cdots$,QS(s=12)} denotes the quantity space of the x-y translation and $QS_{\theta^i_{T_c}}(qp_\theta) \in$ {QS(1),QS(2),$\cdots$,QS(r=16)} denotes the quantity space of orientation.

Finally, an activity should be invariant to the anthropometry of the humans, therefore in order to make the representation scale and orientation invariant, the symbolic representation of all the training data, Υ_j are normalised within the fuzzy qualitative unit circle [-1 1],

$$\begin{cases} QS_{L^i_{T_c}}(qp_l) = qp_l | qp_l \in [\frac{ql_1}{ql}, \frac{ql_2}{ql}, \cdots, \frac{ql_{s-1}}{ql}, 1] \\ QS_{\theta^i_{T_c}}(qp_\theta) = qp_\theta | qp_\theta \in [\frac{q\theta_1}{2\pi}, \frac{q\theta_2}{2\pi}, \cdots, \frac{q\theta_{r-1}}{2\pi}, 1] \end{cases} \tag{4}$$

where x-y translation states qp_l are normalised by the average length of the human body segment *ql* whereas



$$QS_{qp_l}(L^i_{T_c}) = \begin{bmatrix} 0 & 0.0704 & 0 & 0.0141 \\ 0.0845 & 0.1549 & 0.0141 & 0.0141 \\ \vdots & \ddots & \ddots & \vdots \\ 0.8451 & 0.9155 & 0.0141 & 0.0141 \\ 0.9296 & 1.0000 & 0.0141 & 0 \end{bmatrix}$$

$$QS_{qp_\theta}(\theta^i_{T_c}) = \begin{bmatrix} -0.0260 & 0.0260 & 0.0104 & 0.0104 \\ 0.0365 & 0.0885 & 0.0104 & 0.0104 \\ \vdots & \ddots & \ddots & \vdots \\ 0.8490 & 0.9010 & 0.0104 & 0.0104 \\ 0.9115 & 0.9635 & 0.0104 & 0.0104 \end{bmatrix}$$

Fig. 3. Fuzzy qualitative unit circle with resolution s = 16 and r = 12. This shows that the description of the Cartesian translation and orientation in the circle are replaced by quantity space instead. The element of quantity space for every variable in the circle is a finite and convex discretization of the real number line.

the orientation states qp_θ are normalised to 2π, the largest component in the orientation of a unit circle.

The quantisation process which provides a fuzzy qualitative description of its quantitative counterparts within a corresponding region has ruled out the need for cumbersome tracking algorithms. Presently, researchers in this domain had relied on the cumbersome computer vision algorithms or the precision of the sensing devices to obtain the training data. Little work has been conducted to deal with this matter.

C. Constructing the QNTs

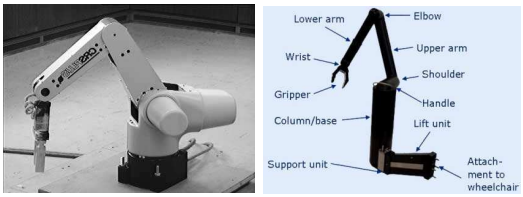
An activity is an ordered combination sequence of all the independent movements performed by the human body segments [2], [8]. Thereby, the objective of this section is to establish an approach to integrate together the motion collected from all the possible joints *i* in the human body during an activity to coherently evaluate the human motion as a whole in image sequences:

$$QNTs_c = \mathbf{A} = \bigoplus_{j=1}^{i} \{\Upsilon^j\} \tag{5}$$

where c is the representation of a typical activity **A** and *i* is the human body points defined in Fig. 2.

In the robotics industry, a multijoint robot manipulator as shown in Fig. 4 is created from a sequence of segment and joint combinations. The segments are the rigid members connecting the joints, or axes. The axes are the movable components of the robot that cause relative motion between adjoining links. In practise, to design a collision free path for multijoint robot manipulator where the motion is constraint by the actuators and its workspace, either forward kinematics or inverse kinematics are employed. In the former, the solution is about finding an end effector's (gripper) pose given a set of joint variables while the latter is to find a set of joint variables that give rise to a particular end effector pose [12].

Exploiting the fundamental theory of the forward kinematics, we parameterised the motion of each body point *i* by six degrees of freedom for the 3D rigid motion. The twist representation has been employed as it provides a more elegant solution and leads to a very simple linear representation of the motion model [12]. Twist ξ is based on



(a) (b)

Fig. 4. An example of a robot manipulator employed in industries.

the observation that every rigid motion can be represented as a rotation around a 3D axis and a translation along this axis. A twist ξ has two representation: a) a 6D vector, or b) a 4 $\times$ 4 matrix with the upper 3 $\times$ 3 component as a skew-symmetric matrix:

$$\xi = \begin{bmatrix} \upsilon_1 \\ \upsilon_2 \\ \upsilon_3 \\ \omega_1 \\ \omega_2 \\ \omega_3 \end{bmatrix} \quad \text{or} \quad \xi = \begin{bmatrix} 0 & -\omega_z & \omega_y & \upsilon_1 \\ \omega_z & 0 & -\omega_x & \upsilon_2 \\ -\omega_y & \omega_x & 0 & \upsilon_3 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{6}$$

ω is a 3D unit vector that points in the direction of the rotation axis. The amount of rotation is specific to a scalar angle θ that is multiplied by the twist: $\xi\theta$ whereas the υ component determines the location of the rotation axis and the amount of translation along this axis (see [12] for a detailed explanation).

In order to realise the motion performed by each limb (hand and leg in this case) in the proposed human model at time T_c, we define the base body reference frame F_0 that is attached to the base body and a spatial reference frame F_a that is static and coincides with F_0 at time T_c. By considering a single kinematics chain of two body parts connected to the base frame, we parameterise the orientation between these connected components in terms of the angle of rotation around the axis of the object coordinate frame θ. This rotation axis in the object frame can be represented by a 3D unit vector ω_1 along the axis, and a point q_1 on the axis. As it is a revolute joint, we can model it by a twist

representation:

$$\xi_1 = \begin{bmatrix} -\omega_1 \times q_1 \\ \omega_1 \end{bmatrix} \tag{7}$$

A rotation of angle θ_1 around this axis can be denoted as:

$$g_1 = e^{\hat{\xi}_1 \theta_1} \tag{8}$$

and the transformation of the point q_1 from F_a coordinates to the base frame $Frame_0$ can be written as:

$$g(\theta_1) = e^{\hat{\xi}_1 \theta_1}.g(0) \tag{9}$$

For a continuous representation from time *t* to time *t+1*, the transformation is:

$$g(\theta_1) = \sum_{1}^{T_c} e^{(\hat{\xi}_1 \theta_1)_{T_c}}.g(0) \tag{10}$$

If there is a kinematics chain of *K* segments where the motion of the K^{th} segment is represented by joint θ_k and each joint is described by a twist ξ_k, the forward kinematics $g_K(\theta_1, \theta_2, \cdots, \theta_k)$ therefore can be computed by the individual twist motion for each joint $e^{\hat{\xi}_k \theta_k}$ and the transformation between the base frame g(0) which is attached to the base body and Frame F_k which is attached to the K segments is:

$$g(\theta_1, \theta_2, \cdots, \theta_k) = e^{\hat{\xi}_1 \theta_1 + \hat{\xi}_2 \theta_2 + \cdots + \hat{\xi}_k \theta_k}.g(0) \tag{11}$$

and the continuous representation is:

$$g(\theta_1, \theta_2, \cdots, \theta_k) = \sum_{1}^{T_c} e^{(\hat{\xi}_1 \theta_1 + \hat{\xi}_2 \theta_2 + \cdots + \hat{\xi}_k \theta_k)_{T_c}}.g(0) \tag{12}$$

In this paper, all the performed activities captured in the video data are fronto-parallel with the camera plane and therefore, only half of the human model is employed to construct the QNTs. We also considered the base body reference frame is located at the hip. Using the achieved normalised symbolic representation from Section II-A and the concept of fuzzy qualitative robot kinematics [6], [9], the product of exponential maps for the arm kinematics chains with respect to the base frame g(0) over a duration T_c is:

$$g_{arm}(QS_{\theta^{1,2,3}}) = \sum_{1}^{T_c} e^{(\hat{\xi}_1 \theta_1 + \hat{\xi}_2 \theta_2 + \hat{\xi}_3 \theta_3)_{T_c}}.g_{arm}(0)$$

where

$$g_{arm}(0) = \begin{bmatrix} I & \begin{bmatrix} L_T \\ L_U + L_L \\ 0 \end{bmatrix} \\ 0 & 1 \end{bmatrix} \tag{14}$$

$$\omega_1 = \omega_2 = \omega_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \tag{14}$$

$$\mathrm{q}_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad \mathrm{q}_2 = \begin{bmatrix} L_T \\ 0 \\ 0 \end{bmatrix} \quad \mathrm{q}_3 = \begin{bmatrix} L_T \\ L_U \\ 0 \end{bmatrix} \tag{14}$$

$$\xi_1 = \begin{bmatrix} -\omega_1 \times q_1 \\ \omega_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \quad \xi_2 = \begin{bmatrix} 0 \\ -L_T \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \quad \xi_3 = \begin{bmatrix} -L_U \\ -L_T \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \tag{14}$$

whereas the product of exponential maps for leg kinematics chains with respect to the same base frame over a duration T_c is

$$g_{leg}(QS_{\theta^{1,2}}) = \sum_{1}^{T_c} e^{(\hat{\xi}_3 \theta_3 + \hat{\xi}_4 \theta_4)_{T_c}}.g_{leg}(0)$$

where

$$g_{leg}(0) = \begin{bmatrix} I & \begin{bmatrix} -L_S - L_V \\ 0 \\ 0 \end{bmatrix} \\ 0 & 1 \end{bmatrix} \tag{15}$$

$$\omega_3 = \omega_4 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \tag{15}$$

$$\mathrm{q}_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad \mathrm{q}_4 = \begin{bmatrix} -L_S \\ 0 \\ 0 \end{bmatrix} \quad \xi_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \quad \xi_4 = \begin{bmatrix} 0 \\ -L_S \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \tag{15}$$

As described earlier, an activity is defined as a combination ordered sequence of all the independent movements performed by the human body segments [2], [8], hence for any given activity, the QNTs is derived as:

$$QNTs_c = g_{arm} \oplus g_{leg} \tag{15}$$

We treat this trajectory of specific state transition pattern in the quantity space alike to the free collision path in robot manipulator workspace as a unique representation of different activities and employed in a classification algorithm. The advantages of the solution are 1) human activity as to activity taxonomy [2], [8] is equivalent to a global motion, that is the union of all of the local motions for the participating body parts over a span of time. Therefore, the choice of robot kinematics which is a well established solution in the robotic community, to describe the motion of the joints of the manipulator and the resulting motion of the rigid bodies which form the robot is utilised, 2) our approach is not a statistical learning method thereby it does not require large training data. Instead strong discriminative features can be derived from just one example activity.

III. EXPERIMENTS

In this section, we present the performance of the proposed approach under different conditions (tracking errors, size of training data and the choice of training data) and a comparison with the Hidden Markov Models (HMMs).

TABLE I

THREE DATA SETS FOR HUMAN MOTION ANALYSIS

Data set	Videos	Subjects	Activities
S1	225	25	3
S2	55	9	6
S3	235	34	5

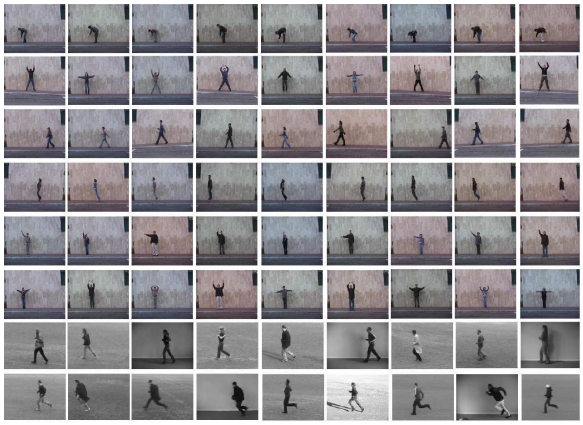


Fig. 5. Sample of the data sets created for the experiments. From top to bottom: bend, jack, walk, jump, wave1, wave2, jog and run, respectively.

A. Data Set

We conducted experiments on two public databases: The **KTH Database** [16] and The **Weizmann Database** [1]. Three data sets were created: 1) **S1**: 225 video streams of 3 activities in 3 planar view scenarios from each of the 25 subjects were selected from the KTH Database. The selected activities were walk, run and jog. The aim is to test the efficacy of the QNTs as walking, running and jogging are activities that exhibit very similar movements but have dramatically different meanings. 2) **S2**: 55 video streams from the six activities were selected from the Weizmann Database. The selected activities were bend, walk, jack, jump, one hand waving (wave1) and two hands waving (wave2). The objective here is to test the effectiveness of the proposed approach in distinguishing a wide variety of activities that are performed by different humans 3) **S3**: All video streams of S1 and only the walking of S2 were selected. The purpose is to test the generality of the QNTs in differentiating the same activities from different environments. The three data sets are summarised in Table I and samples of the data sets are illustrated in Fig. 5.

B. Pre-processing and Training

First of all, we defined a 6 DOF kinematics structure as illustrated in Fig. 2. All joints have an axis orientation parallel to the Z-axis in the camera frame. Then, for each video sequence created in the data sets, the joint track for the five landmarks points on the proposed human model were extracted. After manually initialising the first frame, we employed a tracking approach similar to [7]. The number of particles considered in each level are 2000, 800 and 400 respectively. Then, accordingly to Section II-A these continuous featured motion data are quantised into their associated sequence discrete symbolic representation, *state*. Throughout the experiments, the level of resolution in the fuzzy qualitative unit circle is $s = 12$ and $r = 16$, respectively. Now, we have five time series of state representation per activity. In order to construct the QNTs which are a union of all of the local motions for the participating body parts over a span of time, the qualitative state representations are inputted to the fuzzy qualitative robot kinematics solutions as described in Section II-C. Thus, each activity will be represented as a manifold trajectory of a specific state transition pattern in the quantity space.

C. Results and Analysis

For activity classification, we adopted the nearest-neighbour classifier where the Euclidean metric was used as our distance measure. The recognition results for each data set are shown in Tables II, III and IV, respectively. From the analysis of the results, the following hypotheses can be made:

- For all three data sets, the percentage of correct classification with the proposed approach is beyond all expectation. The mean of classification accuracy for each data set is higher than 80%, even with the case of 400 particles. This has shown that the quantisation process of which continuous dynamic featured data quantised into discrete symbolic representation, state in the fuzzy qualitative unit circle described in Section II-A has successfully bridge the gap between low level processing and high level activity understanding.
- The QNTs are indeed informative as correctly classified human activities to a good extent, in particular in S1 where the three activities exhibit very similar movement. In S1, the QNTs only mis-classified a small number of subjects given by the three tested activities exhibits very similar movement. Further analysis on the mis-classified data found that the activity performed is also barely distinguishable from a human perspective.
- In order to test the robustness of the proposed method, we performed a second set of experiments by selecting a wide range of activities performed by different subjects. These activities are deliberately selected to evaluate the proposed solution. As expected, the successful recognition rate of S2 is perfect as for all the chosen activities, the motions differ greatly from each other (see Table III). For instance, hand waving(wave1 and wave2) is a stationary activity and walking is non stationary horizontal activity.
- From Table IV, it illustrates that the QNTs are generic and insensitive to different motion styles and speeds across different human anatomy. For instance, the constructed walking QNTs from S1 are employed to recognise the walking data in S2 and vice versa. We consider this a satisfactory performance as we were able to maintain the recognition accuracy to a reasonable degree.

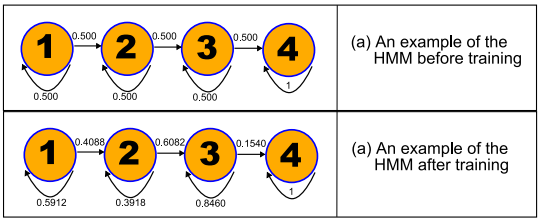


Fig. 6. An example HMMs for an activity

D. Quantitative comparison

A comparison is done with conventional hidden Markov model and our method in classification task. Basically, HMMs are a class of dynamic Bayesian networks where there is a temporal evolution of nodes. A HMMs model λ is specified by the tuple *(Q,O,A,B,π)* where Q is the set of possible state, O is the set of observation symbols, A is the state transition probability matrix $(a_{ij} = P(q_{t+1} = j|q_t = i))$, B is the observation probability distribution $(b_j(k) = P(o_t = k|q_t = j))$ and π is the initial state distribution. It is very straightforward to generalise this model to continuous output models (for more detail, please refer [15]). We choose a four state left-right discrete HMMs for comparison and the pre-processing steps are as [14]. The number of states is empirically determined and it is observed that an increase to a larger number of states did not result in any performance gains in the data sets. Each model (activity) was trained on 1%, 20% and 50% of randomly selected instances of activities and the best (highest-likelihood) models were kept for comparison as HMMs are known to produce models of varying quality, even when trained repeatedly with the same data. An example of the HMMs structure before and after training for an activity is shown in Fig. 6.

The comparison of both classification accuracies are provided in Tables V, VI and VII, respectively. It is observed that on the three data sets, the QNTs perform same/much better than the conventional HMMs. It is worth pointing out that the QNTs employed in this experiment are constructed from 400 particles with 1% training data while the best HMMs were employed for this comparison.

From the analysis of the results, we notice that the effectiveness of the models in HMMs are very much dependant on the accuracy of the training data and the quantity of training data. For instance in Table V, the classification rate of HMMs using the 1%, 20% and 50% training data had a very clear margin whereas the QNTs are fairly constant. The reason is that our solution is not a statistical learning method thereby does not require large training data. Instead strong discriminative features can be derived from just one example activity. A further analysis by employing one subject sequentially as the training data, Table VII shows that the choice of the selection has a huge influence on the recognition rate in HMMs. The worst and the best achieved differs by approximately 44% in the HMMs while the QNTs only differ by approximately 2%. Again, this has proof that the QNTs are generic even with different body anatomy and motion styles. Finally, HMMs are also notoriously sensitive to the precision of featured data. This is notable from Table VI as only 400 particles are employed to predict the featured data and inputted into the HMMs, the mean percentage of successful recognition is only 51% for the three data sets. However at a much higher resolution of the tracking algorithm, the average percentage of successful recognition rate increases to more than 80%. In spite of high recognition of the activity in this case, it should be noted that the number of particles employed in the tracking algorithm is directly proportional to the computational complexity *O(2N)* and time complexity $N \sum_{k=1}^{m} \tau_k$.

TABLE II

RECOGNITION RATE FOR S1

	Particles	Walking	Running	Jogging
S1	(400)	80%	81%	92%
	(800)	82%	81%	92%
	(2000)	80%	81%	93%

TABLE III

RECOGNITION RATE FOR S2

	Particles	Bending	Walking	Jacking	Jumping	One hand waving	Two hands waving
S2	(400)	100%	100%	100%	100%	100%	100%
	(800)	100%	100%	100%	100%	100%	100%
	(2000)	100%	100%	100%	100%	100%	100%

TABLE IV

RECOGNITION RATE FOR S3

	Particles	Walking	Running	Jogging	One hand waving	Two hands waving
S3	(400)	86%(86%)	81%	92%	100%	100%
	(800)	86%(86%)	81%	92%	100%	100%
	(2000)	86%(86%)	81%	92%	100%	100%

TABLE V

COMPARISON WITH HMMS. AVERAGE CLASSIFICATION RATE EMPLOYING DIFFERENT TRAINING DATA SIZES

	HMMs with 1% training data	HMMs with 20% training data	HMMs with 50% training data	Ours 1% training data
S1	54%	75%	77%	**85%**
S2	62%	88%	91%	**100%**
S3	68%	72%	72%	**88%**

IV. CONCLUDING REMARKS

There are always two essential parts in human motion recognition: the low level vision processing and the high

TABLE VI

COMPARISON WITH HMMS. AVERAGE CLASSIFICATION RATE EMPLOYING DIFFERENT TRACKING RESOLUTIONS

	HMMs with featured data 400 particles	HMMs with featured data 800 particles	HMMs with featured data 2000 particles	Ours with featured data 400 particles
S1	38%	78%	82%	**85%**
S2	61%	80%	88%	**100%**
S3	54%	83%	88%	**88%**

TABLE VII

COMPARISON WITH HMMS. AVERAGE CLASSIFICATION RATE WITH USING EACH SUBJECT AS TRAINING DATA ONCE, AND TESTED AGAINST THE REMAINING. THE WORST ACHIEVED ARE IN BRACKETS () AND THE BEST ACHIEVED ARE IN SQUARE []

	HMMs	Ours
S1	54% (38%)[88%]	**85%** (84%)[86%]
S2	75% (64%)[98%]	**100%** (100%)[100%]
S3	67% (32%)[80%]	**88%** (88%)[88%]

level vision understanding that is based on it. In this paper, we have adopted the methodology of fuzzy qualitative unit circle and fuzzy qualitative robot kinematics approaches to model and represent an activity. The advantages are that the quantisation process has enabled us to deal with low precision feature data, therefore avoiding the difficulty of incrementally updating offline learned model which are also time-consuming. Secondly, the exploitation of qualitative robot kinematics which is a well established solution in robotic community to describe the motion of the joints of the manipulator and the resulting motion of the rigid bodies which form the robot. Empirically, we have demonstrated that our method produces a very encouraging recognition rate on two public databases. A comparison with the HMMs also shows that our proposed approach is significant in a variety of aspects.

However, there are several unsolved problems associated with our framework that we are currently investigating. For instance, we are developing alternative methodologies for constructing the unit circle, investigating the best unit circle resolutions in this domain and representing more complicated activities.

REFERENCES

[1] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Action as space-time shapes. In *Proceedings of IEEE International Conference on Computer Vision*, volume 2, pages 1395–1402, Beijing, China, 2005.

[2] A. Bobick and A. Wilson. A state-based technique for the summarization and recognition of gesture. In *Proceeding of the International Conference on Computer Vision*, pages 382–388, 1995.

[3] M. Brand, N. Oliver, and A. Pentland. Coupled hidden markov models for complex action recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 994–999, 1997.

[4] H. Buxton. Learning and understanding dyanmic scene activity. *Image and Vision Computing*, 21(1):125–136, 2003.

[5] L. Campbell and A. Bobick. Recognition of human body motion using phase space constraints. In *Proceedings of the Fifth International Conference on Computer Vision*, pages 624–630, Washington, DC, USA, 1995.

[6] C. S. Chan, H. Liu, and D. Brown. Recognition of human motion from qualitative normalised templates. *Journal of Intelligent and Robotic Systems*, 48(1):79–95, January 2007.

[7] M. Isard and A. Blake. Condensation: Conditional density propagation for visual tracking. *International Journal of Computeter Vision*, 29(1):5–28, 1998.

[8] G. Johansson. Visual motion perception. *Scientific American*, 232(6):76–88, 1975.

[9] H. Liu, D. Brown, and G. Coghill. Fuzzy qualitative robot kinematics (in press). *IEEE Transactions on Fuzzy Systems*, 2007.

[10] H. Liu and G. Coghill. Fuzzy qualitative trigonometry. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, volume II, pages 1291–1296, Hawaii, USA, 2005.

[11] K. Murakami and H. Taguchi. Gesture recognition using recurrent neural networks. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 237–242. ACM Press, 1991.

[12] R. M. Murray, S. Shastry, Z. Li, and S. S. Sastry. *A Mathematical Introduction to Robotic Manipulation*. CRC Press, 1994.

[13] N. Oliver, B. Rosario, and A. Pentland. A bayesian computer system for modeling human interactions. *IEEE Transactions on Pattern Analysis and Machine Intellingence*, 22(8):831–843, 2000.

[14] H. Qiang and C. Debrunner. Individual recognition from periodic activty using hidden markov models. In *Proceedings of the Workshop on Human Motion*, pages 47–52, 2000.

[15] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–284, February 1989.

[16] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *Proceedings of the International Conference on Pattern Recognition*, volume 3, pages 32–36, Hong Kong, 2004.

[17] Q. Shen and R. Leitch. Combining qualitative simulation and fuzzy sets. *Recent advances in qualitative physics*, pages 83–100, 1993.

[18] Q. Shen and R. Leitch. Fuzzy qualitative simulation. *IEEE Transactions on Systems, Man and Cybernetics*, 23(4):1038–1061, Jul/Aug 1993.

[19] H. Sidenbladh, M. Black, and D. Fleet. Stochastic tracking of 3d human figures using 2d image motion. In *Proceedings of the 6th European Conference on Computer Vision*, number II, pages 702–718, London, UK, 2000. Springer-Verlag.

[20] A. Sundaresan, R. Chellappa, and A. Roy Chowdhury. Multiple view tracking of humans modelled by kinematic chains. In *Proceedings of the International Conference on Image Processing*, volume 2, pages 1009–1012, Singapore, 2004.

[21] A. Wilson and A. Bobick. Parametric hidden markov models for gesture recognition. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 21(9):884–900, 1999.

[22] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden markov model. In *Proceedings of the 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 379–385, 1992.