Markerless Multi-view Human Motion Tracking Using Manifold Model Learning by Charting

Sanjay Sainia, Dayang Rohaya Bt Awang Rambli, Suziah Bt Sulaiman, M Nordin B Zakaria, Siti Rohkma

Department of Computer & Information Science
Universiti Teknologi PETRONAS
Bandar Seri Iskandar, 31750, Tronoh, Perak, Malaysia.

Abstract

Computer vision based markerless human motion tracking has gained popularity in various potential application domains including automatic visual surveillance, security, human computer interaction, virtual reality and medical applications. In computer vision tracking, articulated human body is a very challenging issue because of unknown motion types and high dimensionality. The low-dimension approaches have been effective for overcoming the high-dimensionality problem of tracking the various motions. In this paper, we present a manifold motion model learning in low-dimensional subspace using charting, a nonlinear dimension reduction technique which identify and extract the manifold action from the high-dimensional space. We choose the kernel regressor with Relevance Vector Machine (RVM) to construct the interface between action joint configuration and image space (e.g., Silhouette). The proposed framework allows the identification of the learning phase forward and backward mapping. For tracking of all generative components of the framework we proposed the use of Quantum-inspired particle swarm optimization algorithm to handle local minima problem also for providing global optimization results in search space.

Keywords: Motion tracking, manifold charting, Kernel Regressor RVM, Q-PSO.

1. Introduction

Over the last decades, human motion tracking has been gaining popularity due to increasing demand in potential applications including human computer interaction (HCI), smart automatic surveillance, biomedical application (stroke rehabilitation), virtual reality and computer based animated game and movie industries. Most of these applications need accurate and robust tracking.

Human motion tracking is challenging due to the complex and articulate structure of the human body, and especially when a person is not using special clothes or active markers. Currently, there are commercially available vision-based motions tracking systems in the market. For example Xsens MVN, Vicon, and C-motion, which are very popular because these systems provide accurate and reliable results. However, these systems have their drawbacks, which include (i) user needs to wear special costume, (ii) active markers need to be attached to the human body, (iii) more time is required for setting up and (iv) the cost is very expensive. Due to these shortcomings, these systems are not always desirable or even applicable, particularly for certain types of application domains such as medical and sports.
Vision-based markerless human motion tracking technology provides a very attractive solution for the abovementioned deficiencies. Because in markerless human motion tracking technology, users are neither required to wear any special costume, nor any active markers need to be attached to the human body. This technology captures the sequence of images through single or multiple cameras, and estimates the body poses from the observed images [7].

According to current research study human motion tracking can be divided in two main broad classes i.e. discriminative and generative approaches. Discriminative approaches directly map from image space and pose space, whereas generative approaches follow the prediction-match-update principle. This approach predicts the appearance features in switched image then it uses generative framework such as numerical optimization or particle filter. Recently, learning based approaches have been developed which estimate pose directly from state space.

At present, markerless human motion tracking is a very energetic research area in both two dimensions and three dimensions. Recently several authors have developed many methods and algorithms for human motion tracking such as ICP [22], Gaussian Process Latent Variable Model (GPLVM) [20], Relevance Vector Regression [13, 19], Stochastic Meta Descent (SMD) [24], Hierarchical Particle Swarm Optimization (HPSO) [21], Probabilistic Latent Semantic Analysis (PLSA) model [14], and Hierarchical annealed genetic algorithm [23].

Despite the continuous development of markerless human motion technology, several challenges still remain. The present work identifies some major issues. The first issue, which is related to highly articulated structure of the human skeleton, is human structure twist and rotations generate movements at high degree of freedom (25-50 DOF). This gives rise to difficulties when exploring high dimensional state space in practical time [10, 21, 24]. Self-occlusion is a well-known problem in tracking [12, 15, 22, 24]. One of the main sources of the problem is the highly flexible structure of the human body. Thus self-occlusion between body parts is a very common problem due to this partially occluded body parts. As such, tracking can become complicated, and also valuable information about motion is hidden. Another source is the absence of markers; the switched image features are quite noisy due to insufficient image information, low quality image, and also it creates the local minima problem [24]. Other issues are unknown human body movements and shape variations.

In this paper we present a generative approach for effective human motion model learning in low-dimensional subspace using charting. The nonlinear dimension reduction charting technique automatically identifies and extracts the manifold from high-dimensional space, and keeps the intimacy pose in latent space [10]. We propose the forward and backward mapping between image space and joint configuration space. Backward mapping draws the samples better according to recent observation from motion latent space to model appearance. Forward mapping makes the fast construction from joint angle configuration to silhouettes without producing mock image for tracking. Before mapping, our image processing steps improve the image quality as well as extract the image features such as equal size of images, background subtraction and silhouette image creation. Relevance Vector Machine (RVM) is used to construct the mapping between action joint configuration and image space (e.g., Silhouette). For tracking of all generative components of the framework we proposed the use of Quantum-inspired particle swarm optimization algorithm. The proposed work has two main phases; the first learning phase is to learn the motion model in high dimensional space and to estimate the pose in low-dimensional. The second phase is tracking phase. The two phases are described individually in the proposed tracking section.

The remainder of this paper is organized into six sections as follows: an overview of related works is presented in section 2 and in section 3 and 4 proposed methodologies and tracking framework are discussed. Implementation and experimental results are given in section 5 and the paper is concluded in section 6.

2. Related work

The work on human motion tracking has evolved rapidly in the last decades because of great attention given to its applications for human computer interaction (HCI), smart automatic surveillance, biomedical application and virtual reality (VR). Recent human motion tracking survey showed the different analysis and methods of human motion tracking [1, 2, 3]. Several researchers have developed different approaches for 3D model based human motion tracking system which we divided in two categories. One category of approaches employs sampling based tracking techniques. These techniques directly learn the pose in high dimensional pose space. For example, Annealed Gaussian based Particle Swarm Optimization (AGPSO) [7] and Stochastic Meta Descent (SMD) [24]. ICP shape registration is based on 3D human motion tracking by using hierarchical model point selection (HMPS) [22] and hierarchical annealed genetic algorithm [23]. These approaches need good initialization, precise modelling as well as good image features. However sampling based tracking techniques are not applicable for interface in high-dimensional state space.

Another technique is manifold-based learning motion model, which involves transforming high-dimensional observation space into low-dimensional latent space. The latent space can expose the inherent structure of human motion. In contrast, currently model learning based on low dimensional subspace is showing strength in human motion tracking system.
The one main region is high dimensionality and unknown motion types. These approaches try to avoid the need for explicit initializations, and 3D modelling. Learning based approaches directly recover the human poses from low-dimensional space. The low-dimension approaches have been successful at overcoming the high-dimensionality problem while tracking various types of motions.

Currently, several non-linear dimensional reduction techniques have been developed such as local linear embedded (LLE), in which the charts of manifold data is constructed to a single global coordinate system of low-dimensionality while conserving the nearest local neighbour associations [16, 25]. LLE is very less sensitive; small data affect the constructed chart. ISOMAP estimates pair-wise distances in the geodesic space of the manifold, then constructs the data charts between high-dimensional input spaces to low-dimensional coordinate manifold data space using multidimensional scaling (MDS) [26]. However, data set with high curvature, self-intersections and no convex sampling in manifold specs lead both LLE and ISOMAP techniques to failure, and also do not provide global consistence parameterization of input data sets [28]. Both of these techniques suffer the incidence of noisy and sparse data. Some other techniques have been developed such as Kernel principal component analysis (KPCA) [27], manifold charting [10, 11] and Gaussian process latent variable model (GPLVM) [20] which can successful recover the complex motion, but GPLVM get stuck in the local minima [30]. In our proposed work, we have chosen non-linear dimension reduction technique charting in which the high dimensional manifold is represented as a set of overlapping “chart” or “patch”. This technique has shown success at overcoming the abovementioned difficulties. The main aim of charting is to estimate automatically low dimensional parameterization data set which lies on nonlinear manifold in a high dimensional space.

3. Proposed Tracking Framework

Our proposed tracking framework consists of two main phases i.e. first is space learning and second is tracking. During learning phase as illustrated in figure 1, we estimate the pose in low dimensional space and also perform mapping between the silhouette descriptor and latent space pose. While in the second phase, we track the framework components using Quantum-inspired particle swarm optimization technique which provides global coverage as well as global optimization. The mapping between two space using kernel regressor with relevance vector machine provides well interface between high dimension subspace and low-dimension with minimum loss of information.

3.1. Action model learning using charting:

In the learning phase, we estimate the human body using a skeleton model with vector of 25 joint angels. First, we estimate the root joint, then calculate the body joints one by one. The nonlinear dimension reduction charting technique involves several steps.

The data sets are first decomposed into local linear low-dimensional patches before merging into single low-dimensional joint coordinate. Then forward mapping between local joints Ls to Lc and backward mapping from Lc to Ls are established using sparse Bayesian regression such as relevance vector machine (RVM) [17].

We learn the motion model from low-dimensional latent space where several human actions can be represented by a manifold, which can be extracted by the help of charting [10, 11]. Charting establishes non-linear mapping between high-dimensional joint configuration \( R^D \) and low-dimensional \( R^d \).

We define the parameterization mapping function as follows:

\[ f: R^D \rightarrow R^d, \text{ where } d < D \]

Charting involves three main steps namely estimating local linear and intrinsic dimensionality, charting phase, merging and connection phase. Each of the steps is discussed below.
A. Estimating local linear and intrinsic dimensionality: We begin with a given data set \( X = \{x_1, x_2, \ldots, x_n\} \) where \( x_n \in \mathbb{R}^D \) belongs to \( D \)-dimensional sample space, and an estimation that points are samples from a manifold \( M \) of intrinsic dimensionality \( d < D \). Next, we find the nonlinear forward and backward vector space mapping.

**Forward:** \( G(X) \rightarrow Y = \{y_1, y_2, \ldots, y_n\} \), where \( y_n \in \mathbb{R}^d \)

**Backward:** \( G^{-1}(Y) \rightarrow X \)

To estimate local linear scale \( (r) \) and intrinsic dimensionality \( (d) \) according to the growth rate of each point for different values \( r \) are tracked, \( c(r) = \frac{d}{\log n(r)} \log r \)

The maximum value of \( c(r) \) provides the local linear scale \( (r) \) and intrinsic dimensionality \( (d) \) as \( c(r) = 1/d \) linear scale and lower some other scale [11].

B. Charting phase: In charting phase we find the soft partitioning (charts) of the data set under the criteria of each data partitions which have minimum loss of variance between low and high-dimensionality, and maximum connection between subspace points and near neighbourhoods in similar subspace. Here we use two criteria; first the charts are acquired by fitting Gaussian mixture model (GMM) to the data [11]:

\[
p(y_i | \mu, \Sigma) = \sum_j p(y_j | \mu_j, \Sigma_j) p_j = \sum_j N(y_i; \mu_j, \Sigma_j) p_j
\]

Where Gaussian component \( \mu_j \) is denoted by centred local neighbourhoods and axes are defined by eigenvalue of \( \Sigma_j \).

The second criterion indicates that adjacent partitions should have dominant axes that span similar subspaces [11]. This can be prescribed through cross-entropy between GMM models to the two nearest local neighbourhoods [11].

\[
D(N_1 | N_2) = \int dy N(y; \mu_1, \Sigma_1) \log \frac{N(y; \mu_1, \Sigma_1)}{N(y; \mu_2, \Sigma_2)}
\]

C. Merging and connection phase: In connection phase, we build the connection for all specified charts. Here, we follow the methodology similar to Vijay john [10]. We characterise the low-dimensionality \( Z_k \) of the \( k^{th} \) chart acquired by PCA using the orientation frame of first \( d \) eigenvector of the covariance matrix \( \Sigma_k \). The main goal of the connection phase is to merge all charts into a single globally consistent low-dimensional coordinate system. There each chart belongs to a low-dimensional affine transform \( G_k \in \mathbb{R}^{(d+1) \times d} \).

3.2. Image Processing: Image processing plays a very important role in representation of image features in 2D images. Commonly, most researchers extract the image edges and silhouette image as image features. One of the main reasons is simplicity. Obviously, extracting image feature is a part of image processing. In the proposed work, we extract the two major image features i.e. image edge and silhouette image.

3.2.1. Image edge: Image edges are generally used for tracking as they offer well localization for model matching. The image edge provides a good outline for visible body parts which is unaffected by clothing, colour and lighting conditions. Good image feature gives a compact and preferable complete representation of all task-relevant information in the image. In image edge based tracking, a correct image edge provides a more robust tracking because pure silhouette based method is not sufficient for obtaining exact human body pose due to hidden placement and body ambiguity. The colour edge detector can be used to extract additional structural information if the edges are quite well. In our work we use RGB color edge detection method using a mixture of Euclidean Distance between color vectors and vector angle. The details of the image edge step can be found in [4, 31, 32].

3.2.2. GMM based Image descriptor: In general different types of shape descriptors are possible to represent silhouettes such as Fourier descriptor, shape context and Hu moments. In our work, we use the Gaussian mixture model GMM based silhouette image representation which is presented by Feng Guo et al. [5]. The main goal of GMM based descriptor is to reflect a silhouette as a set of intelligible regions in the 2D space like foreground pixel locations which is generated by a GMM. In a given sample silhouette image, the GMM parameters are acquired by the help of expectation–maximization (EM) algorithm, and starting data clustering is represented by K-mean algorithm. The measurement of critical silhouette based on GMM descriptor is similar to that done by the Kullback-Leibler divergence (KLD). KLD identifies the measurement of the distance between two silhouettes (pixel spatial distributions). In our experiment, we used 20 components and one GMM to represent a silhouette.
3.3. Interface or mapping from shape descriptor to pose: In this phase, we build an interface (mapping) between $L_s$ silhouette and human pose from latent space $L_{jc}$ using kernel regressor, which is learned with a RVM [17, 19]. From $L_s$ to $L_{jc}$ is backward mapping and from $L_{jc}$ to $L_s$ is forward mapping. RVM is an approach to sparse Bayesian mixture learning. The main objective of RVM is to find well estimation between subspace joint angle and silhouette for each cluster of Bayesian mixture. The hyper parameters are presented in RVM to control the weights, W. The representation of charting manifold action in low-dimensional space $L_{jc}$ to $L_s$ silhouette with kernel regressor:

$$Y = f_k(y) = W_k\Phi_k(y)$$

Where $Y$ and $y$ original pose, $\Phi_k$ is vector kernel function and $W_k$ sparse matrix weight, which are learned with RVM. In training time, posterior distribution of most number of weight $w_i$ parameters exceeds sparse, peaking around zero. Remaining non-zero weight is considered as the relevance vector. More detail can be found in [17].

4. Tracking

Recently particle swarm optimization algorithm is very popular in human motion tracking because of its simplicity and easy implementation. Previously several version of PSO has been developed to increasing the search speed and efficiency of PSO. Recently developed hierarchical particle swarm optimisation (HPSO) based human motion tracking this system tracking the human body hierarchically [22]. Some of authors have used diverse search constrain to save computation efficiency and some are doing several partition in search space [11, 22]. However these regions may loss some important feature in search space. Therefore these systems are not providing the global optimization and also not giving full accuracy to human motion tracking. Since tracking framework, we focus more in processing time because after an in depth study of the literature we identified that most subspace learning models based on human motion tracking system avoid processing time. In the work presented in this paper, we have solved this problem more thoroughly using Quantum-inspired particle swarm optimization which we have described below.

4.1. Quantum-inspired particle swarm optimization:

The Quantum-inspired particle swarm optimization is based on quantum mechanics and computing intelligence. In Q-PSO model, each particle behaves as a quantum [8, 9]. The Q-PSO searches for global best next pose which is close to latent-space manifold. The propose Q-PSO algorithm taking $M_{best}$ average value of all individual $P_{best}$ particles which is beneficial to handling local minima problem. The process steps of the Q-PSO are described below:

**Step 1-** Initializations: Randomly initialize the population of particle with random position in the D dimensional problem space. $A_i = (P_{i,1}, P_{i,2}, \ldots, P_{i,D})$ where $P_{i,j}$ is the jth dimension of vector $P_i$ which is defined as:

$$P_{i,j}(t) = \Phi_j(t) * P_{best\,i,j}(t) + (1-\Phi_j(t)) * G_{best\,j}(t)$$

Where $j=1, 2\ldots D$; $\Phi_j(t) = [c_1 * r_{1,j}(t)] /[c_1 * r_{1,j}(t) + c_2 * r_{2,j}(t)]$.

Where $r_{1,j}(t)$ and $r_{2,j}(t)$ are random numbers, $c_1$ and $c_2$ are positive constant cognitive and social components respectively:

$M_{best}$ is mean best position that will define the mean value of all the $P_{best}$ particles:

$$M_{best\,j}(t) = (1/N) \sum_{i=1}^{N} P_{best\,i,j}(t)$$

**Step 2-** Evaluate: each particle’s fitness value.

**Step 3-** Associate every particle’s fitness value with $P_{best}$

**Step 4-** Update: the current $G_{best}$ position with $P_{best\,g}(t)$.

**Step 5-** Calculate the $M_{best}$ using equation (4).

**Step 6-** Update the particle’s position: Retaining Monte Carlo method, for updating the particles we apply two conditions. If rand $(0, 1) < P_m$ ($P_m$ is mutation probability) then we will update the particles according to the following equations:

$$a_{i,j}(t+1) = P_{i,j}(t) + \beta * \left| M_{best\,j}(t) - a_{i,j}(t) \right| * \ln (1/\mu), \text{if } k \geq 0.5$$

$$a_{i,j}(t+1) = P_{i,j}(t) - \beta * \left| M_{best\,j}(t) - a_{i,j}(t) \right| * \ln (1/\mu), \text{if } k < 0.5$$
Where $a_{i,j}(t+1)$ position of the $j^{th}$ dimension of $i^{th}$ particle in $t+1$ generation; $\mu, \kappa$ are random numbers distributed uniformly in $[0,1]$ and $\beta$ contraction-expansion coefficients that we used for controlling the convergence speed of the particles.

Otherwise update the particle's position according to:

$$a_{i,j}(t+1) = a_{k,j}(t) + (1-\delta) \ast (a_{l,j}(t) - a_{m,j}(t)) + \delta \ast (G_{best,j}(t) - a_{k,j}(t))$$

Where $\delta = \kappa_{curr} / \kappa_{max}$ and $k, l, m$ are random integer, $i \neq k \neq l \neq m$.

Step 7- Return to step 2 until a stop criterion is met.

5. Implementation and Experimental Results Analysis

A complete system of the proposed framework is implemented in MATLAB using windows7 with 2.0GHz processor. Since the implementation is still at the initial stage, we tested only real image sets captured by 4 TP-link cameras with resolution 640 x 480 at 30 fps. The system was trained and tested for walking sequences and swing hand sequence actions. In this paper, we only reported the subspace learning results.

Figure 3 shows the distance matrices for 153 side view walking silhouette images described in the earlier sections. In order to appraise GMM shape descriptor, we produce distance matrix using Kullback-Leibler divergence (KLD) which measures the distance between two silhouettes. We compared the Fourier descriptor and shape context histogram descriptor [18] and GMM based shape descriptor based on KLD. The qualitative analysis (visual observation) shows that GMM shape descriptors capture the similarity of the image silhouette well. More details on these results can be found in [5, 18].

![Figure 3](image-url)

Figure 4 shows the comparison between charting and other manifold dimension reduction technique. In Isomap and LLE, it is very difficult to classify the manifold action due to the distance of nearest neighbour data is too large. However, in charting we can classify manifold action easily because charting minimizes the cost function. Manifold charting minimizes a convex cost function that measures the quantity of divergence between the linear models on the global coordinates of the data points. The minimizations of this cost function can be accomplished by resolving a generalized eigen problem. The results of charting algorithm show that it will works well in noisy image with the presence of self-occlusion in the image. Figure 5 shows the motion model learning with different walking sequences: graph (a) and (b) shows very noisy environments while graph (c) exhibit less noisy environment thus indicate a good mapping.

![Figure 4](image-url)

![Figure 5](image-url)
6. Conclusion

In this paper, we have presented vision based human motion tracking using non-linear dimension reduction charting technique and Q-PSO optimization algorithm. In this work, the human body structure is extracted using GMM based silhouette descriptor and joint configuration in manifold space belonging to low-dimensional space. For mapping between two spaces we used kernel regressor with relevance vector machine. The dimension reduction technique charting shows the well estimation of dimensionality of the embedded manifold. In this paper only the charting results is presented. The reporting of the results of the Quantum-inspired particle swarm optimization based tracking is in progress. The experimental results show that the charting algorithm works well in noisy image observation as well as with the presence self-occlusion in the image. Future work will include classification of the manifold human action in low dimension subspace in good form.

Acknowledgment

I would like to thank Dr. Dayang, Dr. Suziah Sulaiman, and Dr. Mohamed Nordin Zakaria for the useful discussion on this project. We would also like to thank UTP for providing the Graduate Assistance ship for this project.

References