Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 41 (2008) 2470–2483
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

Generative tracking of 3D human motion by hierarchical annealed genetic algorithm

Xu Zhao*, Yuncai Liu

Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200240, China

Received 29 August 2007; received in revised form 22 November 2007; accepted 3 January 2008

Abstract

We present a generative method for reconstructing 3D human motion from single images and monocular image sequences. Inadequate observation information in monocular images and the complicated nature of human motion make the 3D human pose reconstruction challenging. In order to mine more prior knowledge about human motion, we extract the motion subspace by performing conventional principle component analysis (PCA) on small sample set of motion capture data. In doing so, we also reduce the problem dimensionality so that the generative pose recovering can be performed more effectively. And, the extracted subspace is naturally hierarchical. This allows us to explore the solution space efficiently. We design an annealed genetic algorithm (AGA) and hierarchical annealed genetic algorithm (HAGA) for human motion analysis that searches the optimal solutions by utilizing the hierarchical characteristics of state space. In tracking scenario, we embed the evolutionary mechanism of AGA into the framework of evolution strategy for adapting the local characteristics of fitness function. We adopt the robust shape contexts descriptor to construct the matching function. Our methods are demonstrated in different motion types and different image sequences. Results of human motion estimation show that our novel generative method can achieve viewpoint invariant 3D pose reconstruction.

Keywords: Human motion analysis; Monocular images; Generative model; Evolutionary algorithm; 3D human tracking; Optimal tracking

1. Introduction

The research into capturing 3D human motion from visual cues has received increasing attention in recent years, due to the drive from a wide spectrum of potential applications such as behavior understanding, content-based image retrieval, and visual surveillance. However, although having been attacked by many researchers, this challenging problem is still long standing because of the difficulties conduced mainly by complicated nature of 3D human motion and incomplete information of 2D images for 3D human motion analysis.

In general, tracking 3D human motion from image sequences can be considered as a problem of temporal state estimation while we view the static images situation as the special case of tracking. In the context of graphical models, the state-of-art approaches can be classified as generative and discriminative [1]. *Discriminative approaches* [1–6] try to model the state posterior distribution conditioned on observations directly. The models are constructed usually by finding the direct mappings from observation space $\mathbb{Y}$ (image space) to state space $\mathbb{X}$ (pose space) from the training pairs $\{(\mathbf{x}_i, \mathbf{y}_i)|\mathbf{x}_i \in \mathbb{X}, \mathbf{y}_i \in \mathbb{Y}, i = 1, 2, \ldots, n\}$. Discriminative algorithms allow to fast inference and flexible interpolate in trained regions by absorbing computing expense into the training process. But they may fail on novel inputs, especially if trained using small data sets. Also, accurate learning of one-to-more mapping in observation space is difficult because the conditional state distributions are inherent multimodal. The selection of training samples is also an intractable problem of the approach, which is derived from the difficult tradeoff between generalization capability of the trained model and the training expense. *Generative methods* [7–13] is another typical approach which follows the prediction-match-update philosophy embedded into the framework of bottom-up Bayes' rule. Comparing with the discriminative approach, generative approaches model the state posterior density using observation likelihood or cost function.

* Corresponding author. Tel.: +86 21 34204028; fax: +86 21 34204340.
E-mail addresses: zhaoxu@sjtu.edu.cn (X. Zhao), whomliu@sjtu.edu.cn (Y. Liu).

0031-3203/$30.00
doi:10.1016/j.patcog.2008.01.004

Given an image observation and prior state distribution, the posterior likelihood is usually evaluated using Bayes' rule. This approach has a sound framework of probabilistic support and can achieve significant success for recovering complex unknown motions by utilizing well-defined state constrains. However, generative methods are generally computationally expensive because one has to perform complex search over the state space in order to locate the peaks of the observation likelihood. Moreover, prediction model and initialization are also the bottlenecks of the approach especially in tracking situation.

In this paper, we propose a novel generative approach in the framework of evolutionary computation, by which we try to widen the bottlenecks mentioned above with effective search strategy embedded in the extracted state subspace. Considering the generalization of application scenario, the observation information we utilized comes from an uncalibrated monocular camera. This makes the state estimation get into severe ill-conditioned problem. That is to say, the found solutions could be infeasible even if the search algorithm is powerful enough. The rather that, we have to confront the curse of dimensionality because there are more than 40 degrees of freedom (DOF) of full body joints in our 3D human model. Therefore, the process searching for optimal solutions should be performed in some compact state space by the search algorithms which suit for the characteristics of this space. In doing so, infeasible solutions, namely, the absurd poses can be avoided naturally. To this end, we consider to reduce the dimensionality of state space by principal component analysis (PCA) of motion capture data. Actually, the motion capture data embody the prior knowledge about human motion. By PCA, the aim of both reducing dimensionality and extracting the prior knowledge of human motion are achieved simultaneously. And, from the theoretical view, PCA is optimal in the sense of reconstruction because it allows the minimal information loss in the course of state transformation from the subspace to original state space. Different from the previous works [14,15], we perform the lengthways PCA, by which the subspace can be extracted from only single sequence of motion capture data. Based on theconsistency of human motion, the structure of state subspace is explored with data clustering and thus we can divide the whole motion into several typical phases represented by the cluster centers. The clustering results are used to determine the global rotation of human motion in our algorithm.

To explore the solution space efficiently, we design the annealed genetic algorithm (AGA) combining the ideas of simulated annealing (SA) and genetic algorithm (GA) [16]. In fact, AGA is an evolutionary search strategy built on the base of the evolution of single chromosome ($(1+1)$-ES. Namely, the size of population always is kept as 1.) The convergence of AGA is controlled by some annealing parameters. As the promoted version of AGA, hierarchical annealed genetic algorithm (HAGA) searches the optimal solutions more effectively than AGA by utilizing the characteristics of state space. According to the theory of PCA, in our problem, the first principle component captures the most important part of human motion and the rest of principle components capture the detailed parts of this motion. And, in monocular uncalibrated camera situation, the fitness function (observation likelihood function) is very sensitive to the change of global motions. The HAGA performs hierarchical search automatically in the extracted state subspace by localizing priorly the state variables such as the global motions and the coordinate of the first principle component which dominate the topology of state space. The detailed introduction about both algorithms will be presented in the following sections. The HAGA is used dominantly to estimate human motion from the static images. In tracking situation, we develop the optimal tracking algorithm on the base of $(\mu/\mu, \lambda)$-ES [17] in conjunction with the evolutionary mechanism ofAGA. As for the fitness function, we adopt the shape contexts descriptor [18] to construct the matching function, by which the validity and the robustness of the matching between image features and synthesized model features can be achieved.

1.1. Previous work

There has been considerable previous work on capturing human motion from image information. The earlier work on this research topic had been reviewed comprehensively by the survey papers [19–21]. Generally speaking, to recover 3D human pose configuration, more information are required than image can provide especially in the monocular situation. Therefore, much work focus on using prior knowledge and experiential data in order to alleviate the ill-condition of this problem. Explicit body model embodies the most important prior knowledge about pose configuration and thus be widely used in human motion analysis. Another class of important prior knowledge comes from the experiential data such as motion capture data acquired by commercial motion capture system and some hand-labeled data. The combination of the both prior information can produces favorable techniques for solving this problem.

Agarwal et al. [11] distill prior information (the motion model) of human motion from hand-labeled training sequences using PCA and clustering on the base of a simple 2D human body model. This method presents a good autoregressive-based tracking scheme but has no description about pose initialization. In the framework of generative approach, the prior information is usually employed to constrain or reduce the search space. Urtasun et al. [15,22] construct a differentiable objective function based on the PCA of motion capture data and then find the poses of all frames simultaneous by optimizing a function in low-dim space. Sidenbladh et al. [8,14] present similar methods in the framework of stochastic optimization. For a specific activity, such methods need many example sequences of images to perform PCA, and all of these sequences must keep same length and same phase by interpolating and aligning. Ning et al. [12] learn a motion model from semi-automatically acquired training examples which are aligned with correlation function, and then, some motion constrains are introduced to cut the search space. Unlike these methods, we extract the state subspace from only one example sequence of a specific activity using the lengthways PCA and thus have no use for interpolating or aligning. In addition, useful motion constraints are included naturally in the low-dim subspace.

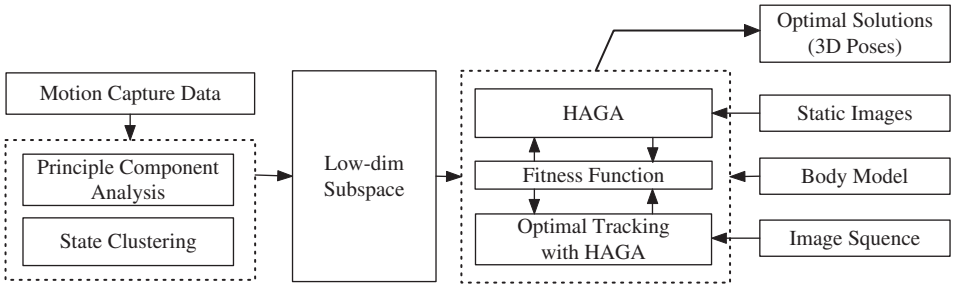

Fig. 1. The framework of our approach.

In recent years, particle filter [23] (also known as condensation algorithm) based optimization methods are used widely for recovering human pose in generative framework [7–13]. However, as a stochastic search algorithm, we think that particle filter is essentially similar with evolutionary algorithm (EA) if having no explicit temporal dynamic model. The EA can provide more flexible evolutionary mechanism such as crossover operator. This is the important motivation for us to solve this problem in the framework of EA. A noticeable example showing the relationship between particle filter and EA is the work of Deutscher et al. [24]. By introducing the crossover operator, the annealed particle filter proposed in their early work [7] get remarkable improvement.

The methods mentioned above utilize the prior information in generative fashion. By contrast, discriminative approaches make use of prior information by learning mapping models directly from training examples. In Ref. [2], Agarwal and Triggs present several regression-based mapping operators using shape context descriptor. The direct prediction of poses from image cues can be achieved using the learned regressor parameters. Sminchisescu et al. [1] learn a multimodal state distribution from the training pairs based on the conditional Bayesian mixture of experts models. In Refs. [3,25], learning specialized nonlinear mappings from Hu moment representation of the input shape and the pose space facilitated successful recovery of the pose directly from the visual input. Elgammal and Lee [5] learn viewbased representations of activity manifolds using nonlinear dimensionality reduction method (LLE). Then, the nonlinear mapping from the embedding space into both visual input space and 3D pose space are learnt using the generalized radial basis function. These methods can bring the interest of fast state inference after finishing the training. However, they are prone to fail when the small training database are used. The styles of using prior information are multiform. Mori and Malik [13] contain the prior information in the stored 2D image exemplars, on which the locations of the body joints are marked manually. By the shape contexts matching with the stored exemplars, the joint positions of the input images are estimated. With this, the 3D poses are reconstructed by the Taylor method [26].

1.2. Framework of our approach

Comparing with the previous methods, extracting the common characteristic of a special types of motion from prior information and represent them with some compact forms are of particular interests to us. At the same time, we ensure the motion individuality of the input sequences with effective evolutionary search strategy suiting for the characteristic of state subspace.

The framework of our approach is illustrated in Fig. 1. We define the state space $\mathbb{X}$, in which a 3D human pose is represented by vector $\mathbf{x}$ corresponding to the motion capture data. From given set consisting of n frames of motion capture data $\{\mathbf{x}_i | \mathbf{x}_i \in \mathbb{X}, i = 1, 2, \ldots, n\}$, we extract the state subspace $\mathbb{X}_s$ by performing conventional PCA. The high-dim state vectors are projected onto the subspace $\mathbb{X}_s$. Then, the k-means clustering is performed in $\mathbb{X}_s$ for human pose estimation. Above steps are introduced in Section 2. In Section 3, we construct the fitness function on the basis of shape contexts descriptor. The silhouettes of people in images are extracted to match with the synthesized model features. Section 4 details the mechanism of our algorithms, AGA and HAGA. How to incorporate the characteristic of the subspace $\mathbb{X}_s$ into the framework of AGA is the core of this section. The capability of HAGA is testified by the experiments where the 3D poses are estimated from static images. In Section 5, we introduce an EA-based optimum tracking algorithm. The initialization of tracking is completed by HAGA. We present experimental results on different motion type sequences. In both static and tracking situation, the optimization process is performed in low-dim space. The final 3D poses are outputted by the PCA injection. Section 6 concludes with brief summary, some discussions and directions for future work.

2. State space analysis

The state space $\mathbb{X}$ contains all of the legal and illegal 3D poses, in correspondence with real human motion, represented by the joint angles vectors of body model. The potential special interests motivate us to analyze the characteristics and structure of this space. Such interests involve mainly modeling the human activities effectively in the extracted state subspace and eliminating the curse of dimension.

2.1. Pose representation

We use a explicit model that represent the articulated structure of the human body. Our fundamental 3D skeleton model

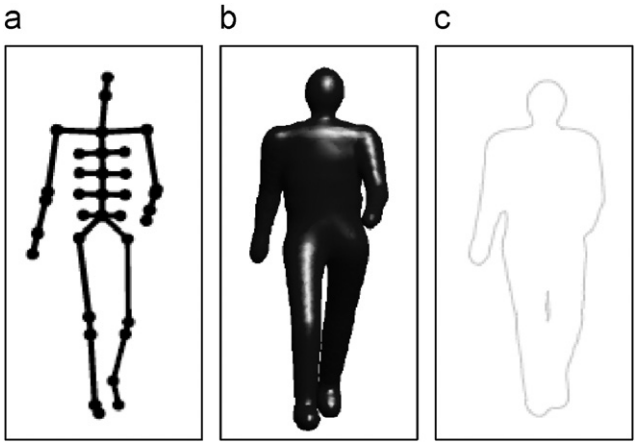

Fig. 2. (a) The 3D human skeleton model. (b) The 3D human convolution surface model. (c) The 2D convolution curves.

(see Fig. 2a) is composed of 34 articulated rigid sticks. The pose is described by a 44D vector $\mathbf{x} = \{\mathbf{x}_g, \mathbf{x}_j\}$, where 3D vector $\mathbf{x}_g$ represents the global rotations of human motion and 41D vector $\mathbf{x}_j$ represents the joint angles.

Fig. 2b shows the 3D convolution surface [27] human model which actually is an isosurface in a scalar field defined by convolving the 3D body skeleton with a kernel function [28]. Similarly, the 2D convolution curves of human body as shown in Fig. 2c are the isocurves generated by convolving the 2D projection skeleton. As the synthetical model features, the curves are used to match with the edges of image silhouettes for constructing the likelihood function.

2.2. Extracting and analyzing the subspace

All of the 3D poses distribute in the state space $\mathbb{X}$. The pose set that belongs to a special activity, such as walking, running, handshaking, etc., generally crowd in a subspace of $\mathbb{X}$. We extract the subspace $\mathbb{X}_s$ from motion capture data obtained from the CMU database (http://mocap.cs.cmu.edu/).

Assuming $\{\mathbf{x}_t | \mathbf{x}_t \in \mathbb{X}\}$ is a given sequence of motion capture data corresponding to one motion type, where t is the time tag, the subspace $\mathbb{X}_s$ is extracted by PCA as follows:

(1) Centering the state vectors and assembling them into a matrix (by columns): $\mathbf{X} = [(\mathbf{x}_1 - \mathbf{c})(\mathbf{x}_2 - \mathbf{c}) \cdots (\mathbf{x}_T - \mathbf{c})]$, where $\mathbf{c}$ is the mean vector.
(2) Performing a singular value decomposition of the matrix to project out the dominant directions: $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}}$.[1]
(3) Projecting the state vectors into the dominant subspace: each state vector is represented as a reduced vector $\mathbf{x}_s = \mathbf{U}_{\mathbf{m}}^{\mathrm{T}}(\mathbf{x} - \mathbf{c})$, where $\mathbf{U}_m$ is the matrix consisting of first m columns of $\mathbf{U}$, by which the m–D subspace $\mathbb{X}_s$ is spanned.

[1] PCA in this way is equivalent to that from a covariance matrix because the left singular vectors of $\mathbf{X}$ are same as the eigenvectors of matrix $\mathbf{X}\mathbf{X}^{\mathrm{T}}$.

Table 1
The cumulative sum of principal component variance percentage

Motion type	Cumulative sum of principal component variance percentage (%) (the first five bases)				
Walking	58.69	79.18	88.67	94.50	96.02
Running	54.85	77.18	93.01	95.21	96.95
Handshaking	51.04	69.09	81.17	85.94	89.12

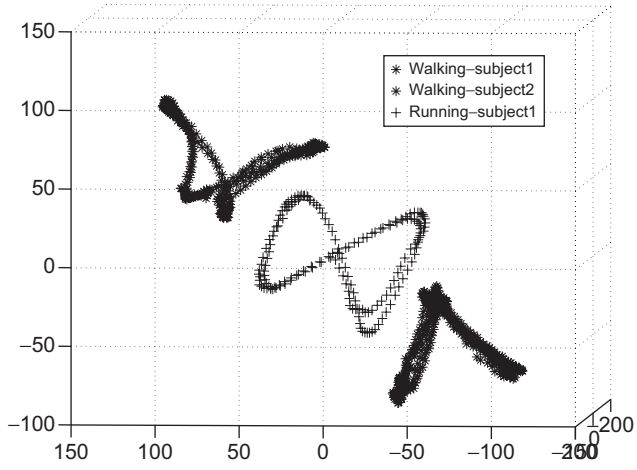

Fig. 3. The manifolds of walking sequences and running sequence in 3D subspace. The bases of the extracted subspace construct the coordinate axes.

Therefore, the original state vector $\mathbf{x}$ can be reconstructed by

$$\mathbf{x} = \mathbf{c} + \mathbf{U}_m \mathbf{x}_s. \quad (1)$$

The dimensionality m of subspace $\mathbb{X}_s$ is determined according to the cumulative sum ε of principal component variance percentage. With our experiences, the value of ε is set to be not smaller than 0.95; accordingly, the value of m is not greater than 6 generally. It means that the PCA injection from $\mathbf{x}_s$ only lose negligible information. This can be seen from Table 1.

In this way, we extract the subspace $\mathbb{X}_s$ of one type of human motion from single training sequence. Actually, similar low-dim subspace can be extracted from the training sequences that belong to the same type of motions but performed by different subjects. And, the training sequences corresponding to different type of motions produce different subspace. For example, experiments demonstrate that different walking sequences generate similar manifolds in the 3D subspace, which is different from that of running motion. See Fig. 3.

In subspace $\mathbb{X}_s$, the special human motion shows the special manifold structure which indicates the common identity of the type of motion. Based on the consistency of human motion, we partition the manifolds into different subparts with the k-means clustering and each subpart represents different phase of human motion. Here, we choose the number of clustering to be four and represent the four clustering centers as $\mathbf{x}_{c_1}, \mathbf{x}_{c_2}, \mathbf{x}_{c_3}, \mathbf{x}_{c_4}$, respectively. Fig. 4 shows the clustering outcome in $\mathbb{X}_s$ and the corresponding joint angles. Actually, the

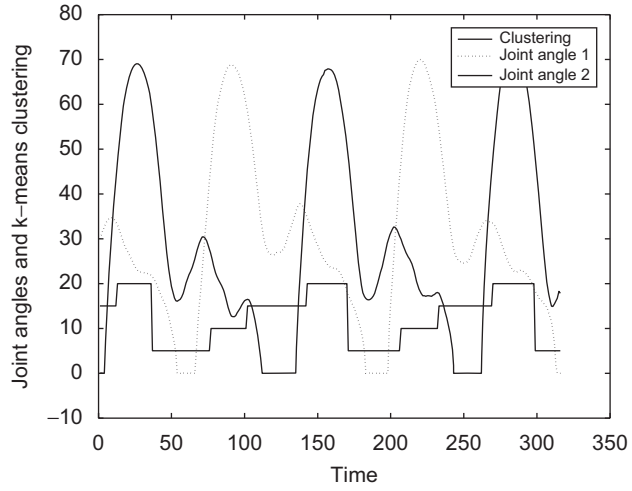

Fig. 4. The k-means clustering of low-dim state vectors and the corresponding joint angles. The stepwise lines represent the clustering labels.

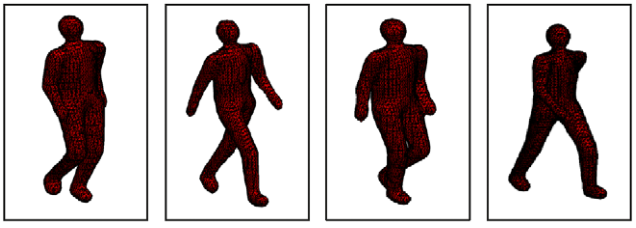

Fig. 5. The 3D human poses correspond to the clustering centers in low-dim subspace $\mathbb{X}_s$, which actually are the key frames of a walking sequence.

clustering centers correspond to the key frames of the motion sequence because they are also the centers of the special motion phases. This can be seen from Fig. 5, in which the clustering centers of a walking sequence described by 3D human poses are illustrated.

3. Fitness function

In generative framework, pose capturing can be formulated as Bayesian posterior distribution inference:

$$p(\mathbf{x}_s|\mathbf{y}) \propto p(\mathbf{x}_s)p(\mathbf{y}|\mathbf{x}_s), \tag{2}$$

where $\mathbf{x}_s$ indicates that the optimal solutions are searched in state subspace $\mathbb{X}_s$. The function $p(\mathbf{y}|\mathbf{x}_s)$ represents the likelihood observing in image $\mathbf{y}$, conditioned on a pose candidate $\mathbf{x}_s$. It is used to evaluate every pose candidate generated from $p(\mathbf{x}_s)$ (in our algorithm, the AGA and HAGA). In the context of EA, the likelihood function is just the fitness function. This function is crucial for pose estimation because as the interface between practical problem and search algorithm, the fitness function influence the validity of the found solutions and the search efficiency to a large extent. We propose a fitness function on the basis of shape contexts matching [18].

The problem of capturing human motion from images requires a robust, discriminative representation of image observation. We choose the image silhouette of subject as the observed image feature, which is extracted using statistical background subtraction. In Fig. 6, we illustrate the process of extracting the image feature. The shape context descriptor is used to describe the shape of image silhouette and convolution curves generated by the pose candidate (see Fig. 2). Fig. 7 illustrate the shape contexts [18] (histograms of local edge pixels into log-polar bins) of human shape. Our shape contexts contain 12 angular × 5 radial bins, giving rise to 60D histograms as shown in Fig. 7b. In the matching process, the regularly spaced points on the edge of the silhouette are sampled as the query shape. The point set sampled from the convolution curves is viewed as the candidate shape. In the experiments, we sample 100 points from image edges and model curves, respectively. Before matching, the image shape and the candidate shape are normalized to same scale. We represent the query shape and the candidate shape as $S_{query}(\mathbf{y})$ and $S_m(\mathbf{x}_s)$, respectively. To this end, the matching cost function is formulated as

$$F(S_{query}(\mathbf{y}), S_m(\mathbf{x}_s)) = \sum_{j=1}^{r} \chi^2(H^j_{query}(\mathbf{y}), H_m(\mathbf{x}_s)^*), \tag{3}$$

where H is the shape context, r is the number of sample point on the edge of image silhouette, and $H_m(\mathbf{x}_s)^* =$

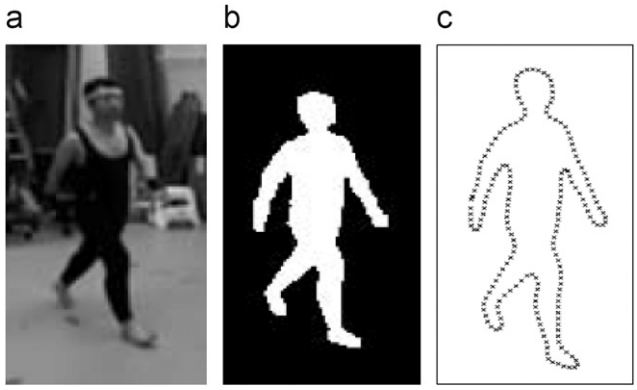

Fig. 6. (a) Original image. (b) Image silhouette extracted by background subtraction. (c) The sampled points on the edge of the silhouette.

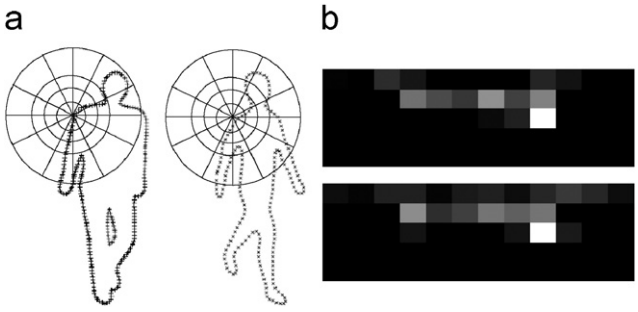

Fig. 7. (a) The shape contexts computed from edge points of image silhouette (right) and sampled points of convolution curves (left). (b) The example shape contexts for reference samples showed in (a) of image silhouette (bottom) and convolution curves (top).

$\arg\min_u \chi^2(H^j_{query}(\mathbf{y}), H^u_m(\mathbf{x}_s))$. Here, we use the χ^2 distance as the similarity measurement. The value of matching cost function $F(S_{query}(\mathbf{y}), S_m(\mathbf{x}_s))$ denotes the extent of similarity between query shape and candidate shape. In our problem, we wish to find the optimal solutions corresponding to the minimal matching cost by searching the state subspace $\mathbb{X}_s$ with the optimization algorithm. In AGA, the optimization mechanism are designed for searching the maximal value of object function. Therefore, according to Eq. (3), the fitness function can be formulated as

$$\mathscr{F}(S_{query}(\mathbf{y}), S_m(\mathbf{x}_s)) = C \cdot \exp(-F(S_{query}(\mathbf{y}), S_m(\mathbf{x}_s))), \quad (4)$$

where C is a constant for adjusting the value range of fitness function.

4. Pose estimation from static images

In this section, we describe the key algorithms of the generative framework, namely, the AGA and HAGA, and their adaption for pose capturing from static images. For clarity, we redefine the full 3D pose vector as $\mathbf{x} = \{\mathbf{x}_g, \mathbf{x}_s\}$, where $\mathbf{x}_g$ is the global motion of human body with respect to the camera and $\mathbf{x}_s$ is the pose vector in state subspace. We perform the state posterior inference by optimizing the fitness function (see Eq. (4)). The optimal pose can be represented as

$$\mathbf{x} = \arg\max_{\mathbf{x}} \mathscr{F}(\mathbf{y}, \mathbf{x}). \quad (5)$$

We maximize the search efficiency by embedding the global search capability of HAGA into the local conditions of state subspace.

4.1. Annealed genetic algorithm

Combining SA and GA, we design the AGA, which actually is a hybrid $(1+1)$ evolutionary strategy. The fundamental idea of SA is to allow moves resulting in solutions of worse quality than the current solution (uphill moves) in order to escape from local minima. The evolution of system state is controlled by the termination condition and stop criteria. GA gains inspirations from the language of natural genetics and biological evolution. The capability searching for global optimal solutions in parallel is the most attractive advantage of GA. Detailed introduction about SA and GA can be found in Ref. [16].

In our algorithm, the local optimal solutions are avoided by introducing several genetic evolutionary principles. We employ the mechanisms which is analogous to the termination condition and stop criteria in SA to control the evolutionary process but not set explicit temperature parameters. We represent the chromosome as $\mathbf{z} = [z_1, z_2, \ldots, z_n]$, where the genes $\{z_i \mid i = 1, 2, \ldots, n\}$ are random numbers uniformly distributed in the interval (0, 1) and n is the dimensionality of state vector. For our problem, each gene of the chromosome corresponds to a component of the pose vector. Here, we use real encodings. The algorithm searching for optimal solutions with the AGA is described as follows:

Table 2
The genetic operators in AGA

Operators	Example
Exchange	$\mathbf{z} = [z_1, z_2, z_3, z_4, z_5, z_6] \longrightarrow \mathbf{z}' = [z_1, z_6, z_3, z_4, z_5, z_2]$
Segment reversion	$\mathbf{z} = [z_1, z_2, z_3, z_4, z_5, z_6] \longrightarrow \mathbf{z}' = [z_1, z_6, z_5, z_4, z_3, z_2]$
Segment shift	$\mathbf{z} = [z_1, z_2, z_3, z_4, z_5, z_6] \longrightarrow \mathbf{z}' = [z_1, z_6, z_2, z_3, z_4, z_5]$
Point mutation	$\mathbf{z} = [z_1, z_2, z_3, z_4, z_5, z_6] \longrightarrow \mathbf{z}' = [z_1, z_2, z'_3, z_4, z_5, z_6]$
Segment mutation	$\mathbf{z} = [z_1, z_2, z_3, z_4, z_5, z_6] \longrightarrow \mathbf{z}' = [z_1, z'_2, z'_3, z'_4, z'_5, z'_6]$

Parameter initialization set values for evolution control parameters.
S_t—stop criteria;
N_t—termination condition;
E_t—times for searching a equation state;
for $s_t = 1$ **to** S_t **do**:
  $NonImproveNum \leftarrow 0$;
  Generate the genes of $\mathbf{z}$ uniformly at random in the interval (0, 1);
  Evaluate the fitness function $\mathscr{F}(\mathbf{z})$ by mapping $\mathbf{z}$ into the problem domain;
  while ($NonImproveNum < N_t$) **do**
    for $e_t = 1$ **to** E_t **do**:
      Evolution of $\mathbf{z}$ driven by the genetic operators; (see Table 2)
      Evaluate $\mathscr{F}(\mathbf{z})$;
    end for
    If the value of fitness function is improved,
    $NonImproveNum \leftarrow 0$, else
    $NonImproveNum \leftarrow NonImproveNum + 1$;
  end while
  Record the optimal $\mathbf{z}$;
end for

We design five genetic operators, which are executed orderly in AGA. We introduce the operators by evolving a example chromosome $\mathbf{z} = [z_1, z_2, z_3, z_4, z_5, z_6]$. The new chromosome generated by the operators is denoted as $\mathbf{z}'$. Assuming the positions generated randomly are numbers 2 and 6 or 3 (for point mutation operator), for example, the five operators are illustrated in Table 2. (The new genes are represented as z'.) The application order of the genetic operators in the algorithm just is as that listed in Table 2.

4.2. Hierarchical annealed genetic algorithm

An effective method to reduce computational efforts required in searching a high dimension space is state space decomposition. In practice, some components of state vector play more important roles than others. Accordingly, the fitness function is more sensitive to these components. Partitioning the state space into several sections according to the "importance" of state components can reduce the cost of searching the space to one that increases linearly with the number of partitions instead of one that increases exponentially with the number of state space dimension.

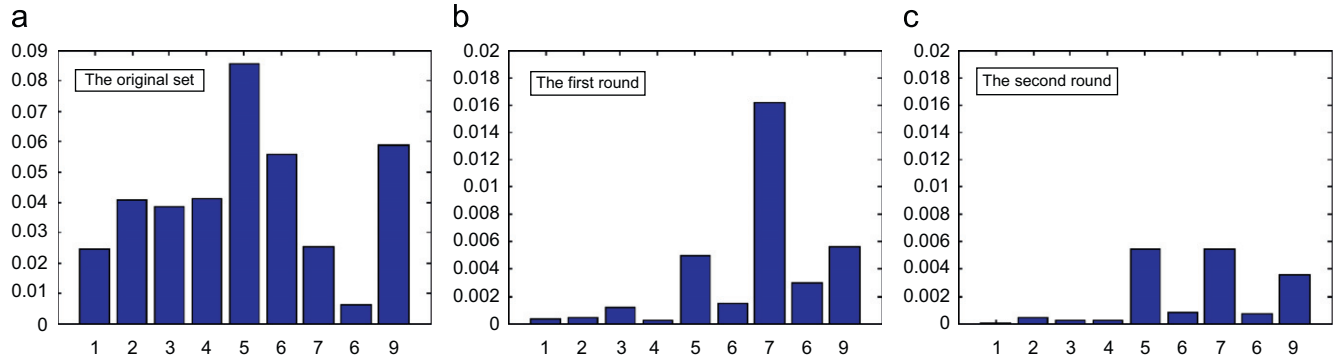

Fig. 8. Variance reduction contrast between principal state components and other state components. Graph (a) shows the variances of state set in which the chromosomes have not been evolved, displaying almost equal variances for each components. Graph (b) shows the variances of state set which have come through one round of state evolution, noticing that the variances of first four principal state components have been greatly reduced whereas the variances of other components have been reduced with a slighter extent. In graph (c), the variances of the principal components have been reduced to very small values indicating advanced localization after coming through two rounds of state evolution.

On the basis of AGA, we develop a HAGA by utilizing the characteristics of state space $\mathbb{X}$. In our model, the use of PCA produces naturally the hierarchical state space. Both soft partition and hard partition of $\mathbb{X}$ can be incorporated into the framework of HAGA. The concept of soft partition in HAGA is similar to that demonstrated in Ref. [24]. Under the soft partition, one need not know which state components are more important in advance. The state space is decomposed automatically by computing the variances of state components which are generated in each annealing run. The smaller variances correspond to the state components that have more influence on the fitness function. Therefore, according to the variances of state components, the state space is partitioned by localizing down the important components to a small area in their range. It is explainable in theory because the important state components dominate the topology of the state space and the little changes of their value can produce great effect whereas the values of other state components had little influence on whether they were selected or not. The theory of soft partition is illustrated in Fig. 8. Comparing with the soft partition, the hard partition decomposes state space in a more direct way, where the topological dominance of state components are known beforehand. In our work, the state space is decomposed by soft partition.

The detailed description of HAGA under soft partition is presented as follows. Because the framework of HAGA is identical with that of AGA, we focus only one annealing run of sate evolution ($s_t \rightarrow s_t + 1$). Preserving the symbol system of AGA, each round of state evolution can be broken down as follows:

(1) Generate initial chromosome $\mathbf{z}=[z_1, z_2, \ldots, z_n]$ at random, where $\{z_i \mid i = 1, 2, \ldots, n\}$ are random numbers uniformly distributed in the interval $(0, 1)$. Map it linearly into the variance domain:

$$\mathbf{z} \mapsto \mathbf{z}_t \in (\min\ \mathbf{z}_t, \max\ \mathbf{z}_t). \tag{6}$$

In the first round of state evolution, $(\min\ \mathbf{z}_1, \max\ \mathbf{z}_1) = (\mathbf{0}, \mathbf{1})$, where t is the mark of state evolution round. By mapping $\mathbf{z}_t$ into the problem domain, the fitness function $\mathscr{F}(\mathbf{z})$ is evaluated.

(2) Evolve the chromosome according to the state evolutionary mechanism of AGA. Before evaluating the fitness function, every new chromosome needs to be mapped into the variance domain as formulated in Eq. (6).

(3) Store N best chromosomes and computing the covariance matrix:

$$\mathbf{V}_{t+1} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{z}_{t+1}^{i} - \mathbf{z}_{t+1}^{c})(\mathbf{z}_{t+1}^{i} - \mathbf{z}_{t+1}^{c})^{\mathrm{T}}, \tag{7}$$

where $\mathbf{z}_{t+1}^{c}$ is the mean vector, and the covariance matrix $\mathbf{V}_{t+1}$ is a diagonal matrix on the assumption that the state components are independent each other. To this end, the variance domain can be formulated as

$$\begin{cases} \min\ \mathbf{z}_{t+1} = \mathbf{z}_{t+1}^{c} - \mathbf{V}_{t+1}\mathbf{c}_{t+1}, \\ \max\ \mathbf{z}_{t+1} = \mathbf{z}_{t+1}^{c} + \mathbf{V}_{t+1}\mathbf{c}_{t+1}, \end{cases} \tag{8}$$

where $\mathbf{c}_{t+1} = [c_{t+1}, c_{t+1}, \ldots, c_{t+1}]$ is used to adjust the variance domain and c_{t+1} is a positive constant.

(4) The variance domain $(\min\ \mathbf{z}_{t+1}, \max\ \mathbf{z}_{t+1})$ is used to cut down the state space in the next round of state evolution.

4.3. Experiments

In this section, we describe the adaption of HAGA for pose estimation from static images. The results are provided with various experiments to test the effectiveness of the proposed generative algorithm in generalizing to different types of motions and different image sequences. For each type of motion, we distill the general prior information about this motion from only one sequence of motion capture data. The translation motion vector is discarded because it is inessential in our model. As for the image feature, we focus only on the shape topology described by the shape contexts.

4.3.1. Global motion

The global motion of human body is very important for its visual appearance in an image and is also critical in disambiguating the left–right confusion. Determining this motion accurately makes our method being viewpoint invariant. In state

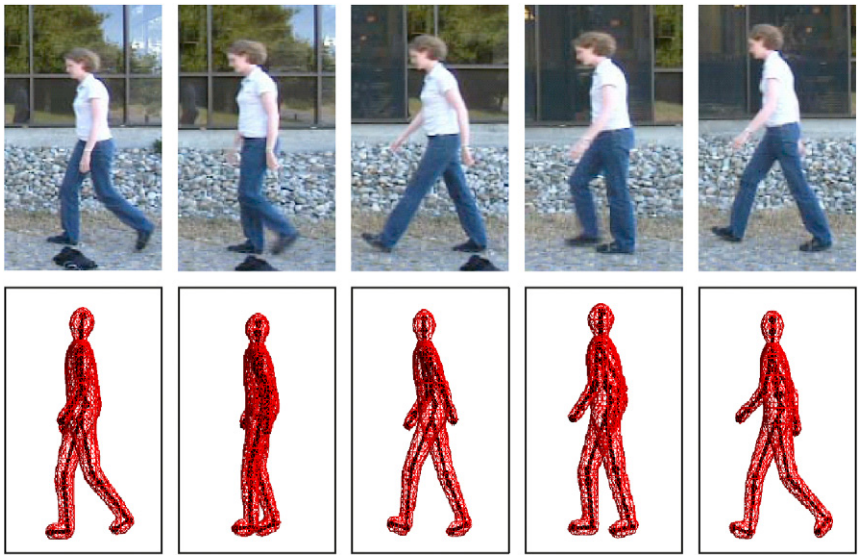

Fig. 9. Results of recovering the poses of a subject walking straight (the images are part of a sequence from http://www.csc.kth.se/~hedvig/data.html). The top row shows the original images and the bottom row shows the reconstructed 3D poses. The second pose demonstrated the left–right confusion in the silhouette.

vector $\mathbf{x}=\{\mathbf{x}_g, \mathbf{x}_s\}$, the global motion $\mathbf{x}_g=\{x_{rx}, x_{ry}, x_{rz}\}$ include the rotation of the full body about the coordinate axes X, Y, Z, respectively. With the aim of both cutting the search space and determining the motion direction roughly, we design the following computation steps that can be incorporated into the framework of HAGA.

(1) In the first round of state evolution ($s_t = 1$), we only actually search the optimal solutions of global motion. Other state components of $\mathbf{x}$ are taken as one of the clustering centers $\mathbf{x}_{c_1}, \mathbf{x}_{c_2}, \mathbf{x}_{c_3}, \mathbf{x}_{c_4}$ randomly. The variance domain (min $\mathbf{x}_g$, max $\mathbf{x}_g$) of $\mathbf{x}_g$ is computed by storing the N best chromosomes. N is determined empirically according to the threshold value of fitness function.
(2) In the rest rounds of state evolution, the chromosome is evolved normally as described in Section 4.2.

In doing so, we can get the coarse scopes of global motion in the first round of state evolution and the fine tuning of these parameters can be achieved in the followed evolution rounds.

4.3.2. Walking motion: straight walk and turning walk

To extract the motion subspace of walking, a data set consisting of motion capture data of a single subject was used. The total number of 316 frames was used. It was found that the different subject and different frame numbers can produce generally identical subspace. To keep the ratio of information loss lower than 0.05, the dimensionality of the subspace was choose to be 5. We test the algorithm in two image sequences, including one straight walk sequence and one turning walk sequence. The purpose of the experiments is to test the capability of the method to cope with limb occlusion and left–right ambiguity.

For the sequence of one subject walking in a straight line, the parameters of HAGA are set as $S_t = 2$, $N_t = 2$, $E_t = 5$. The results are shown in Fig. 9. It can be seen that the estimator is successful in determining the correct global motion as well as the 3D pose of the subject. The occlusion problem are tackled by searching the optimal pose in the extracted subspace because the prior knowledge about walking motion is contained in this space. The left–right confusion is mostly disambiguated because of the special step for searching the global motion. However, in few frames, the left–right confusion conduced by silhouette ambiguity still exist. This can be seen from Fig. 9.

The second sequence tests the capability of generalization of our method in estimating the 3D pose of turning walk. In this sequence [29], a subject is performing a continuous turning walk around a circle, therefore the global motion is changed in a wide range. We found that setting the parameters of HAGA as $S_t = 3$, $N_t = 2$, $E_t = 5$ is adequate for estimating this motion. The results are shown in Fig. 10. We note that the estimator is sensitive to the change of global motion and the special steps in searching for global motion play a important role in the process of pose reconstruction.

4.3.3. Running motion

We will demonstrate that our algorithm is efficient for running motion. According to the reconstruction framework, the algorithm can be generalized to any other types of motions as long as the corresponding subspace can be properly extracted from training data. We extend the types of motion to running motion to test the efficiency of the estimator. The subspace of running motion is extracted from motion capture data consisting of 130 frames. This subspace extracted is more compact than that of walking motion. To keep the ratio of information loss lower than 0.05, it is enough to set the dimensionality of the subspace to be 4.

In the test image sequence, a subject is performing running motion toward the camera therefore the scale of

Fig. 10. Results of recovering the poses of a subject performing a turning walking motion.

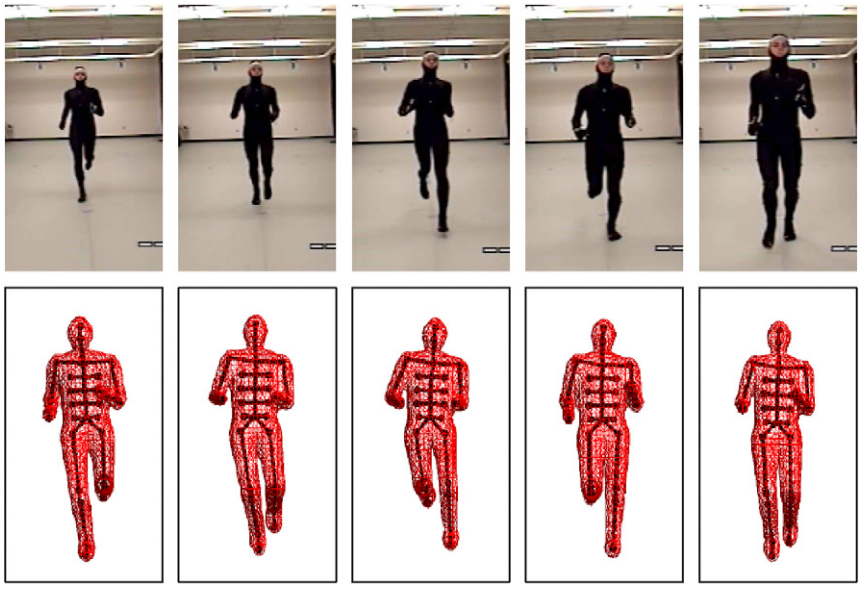

Fig. 11. Results of recovering the poses of a subject performing a running motion. The images are extracted from the video taken from the web site http://mocap.cs.cmu.edu/.

subject is gradually bigger. The scale invariance of the shape context descriptor ensures that the pose estimation is independent on the change of scale. We set the parameters of HAGA as $S_t = 2$, $N_t = 2$, $E_t = 3$. The results can be seen in Fig. 11.

4.3.4. Performance

We conducted performance analysis by recovering the poses in simulated images that are synthesized by our body model (see Section 2). The motion capture data is viewed as ground truth, by which the body model is activated. To evaluate the results of pose recovering, we use the evaluation metrics introduced in Ref. [29]. The average error over all joint angles (in degrees) is defined as

$$D(\mathbf{x}, \hat{\mathbf{x}}) = \sum_{m=1}^{M} \frac{\|x_m - \hat{x}_m\|}{M}, \tag{9}$$

where $\mathbf{x} = \{x_1, x_2, \ldots, x_M\}$ and $\hat{\mathbf{x}} = \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_M\}$ are ground truth pose and estimated pose, respectively. For the sequence of T frames, the average performance and the standard deviation of the performance are computed using the

Table 3
Ground truth and estimated results of some joint angles for walking and running motion

Joint	Walking		Running	
	Ground truth	Estimated value	Ground truth	Estimated value
LFemur	(−35.8261, 10.5333, −18.9678)	(−31.9459, 8.9614, −16.0432)	(−10.2279, 0.2993, −26.6616)	(−14.0521, 2.7866, −21.3780)
RFemur	(−6.2589, −1.4334, 31.2542)	(−3.2185, −4.0614, −31.0695)	(−30.1766, −6.5652, 17.7752)	(−23.4473, −3.8051, 23.2758)
LKnee	(58.2190, 0, 0)	(55.4119, 0, 0)	(42.1629, 0, 0)	(46.5631, 0, 0)
RKnee	(21.5282, 0, 0)	(25.2662, 0, 0)	(104.6430, 0, 0)	(99.0714, 0, 0)
LHumerus	(−36.0573, 10.4688, 81.2025)	(−38.4567, 5.5417, 83.8051)	(−14.8225, −20.1244, 85.3866)	(−18.1869, −16.7721, 87.4147)
RHumerus	(−31.6909, 19.3908, −86.8909)	(−31.3147, 12.6874, −82.4944)	(−54.1141, 8.7512, −89.8125)	(−50.6856, 11.7029, −85.4006)

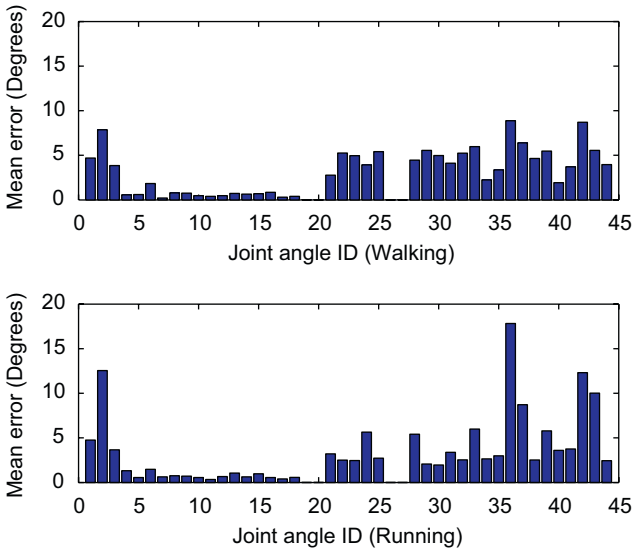

Fig. 12. Mean error of individual joint angle.

following [29]:

$$\mu_{seq} = \frac{1}{T}\sum_{t=1}^{T} D(\mathbf{x_t}, \hat{\mathbf{x}}_\mathbf{t}), \tag{10}$$

$$\sigma_{seq} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}[D(\mathbf{x_t}, \hat{\mathbf{x}}_\mathbf{t}) - \mu_{seq}]^2}. \tag{11}$$

Table 3 shows the ground truth and estimated values of some joint angles in a example frame. Three values in each cell are the rotation angles of the joints around X, Y, Z axes, respectively. The values come from a frame on the level of average error. Actually, other frames show generally the similar comparison results. We also reported the mean errors for each individual joint angle over all test frames, which are shown in Fig. 12. The mean errors of some joint angles are more larger than others because they have more wider range of variation or less observability related to 2D image features. Our results are competitive with others reported in the related literatures. However, the pose estimation still suffers from unsmoothed temporal transfer, which can be reduced by utilizing contextual observation information.

5. Pose tracking from image sequences

We have reconstructed 3D human motion from static images where the pose estimation problem is of static nature. However, in most of the cases, recovering 3D human pose from image sequences can be viewed as a problem of temporal inference with dynamic nature and should be solved in tracking framework. In tracking situation, the previous estimation results can be used to cut the current search space. And, for our problem, the usage of previous observation information is advantageous for disambiguate the left–right confusion shown in Fig. 9. From the Bayes' view, we can formulate the pose tracking problem as

$$p(\mathbf{x}_t|\mathbf{y}_t) \propto p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{x}_{t-1}), \tag{12}$$

where $\{\mathbf{x}_t, t = 1, 2, \ldots, T\}$ and $\{\mathbf{y}_t, t = 1, 2, \ldots, T\}$ represent temporal states and observations, respectively. How to determine the conditional distribution $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ effectively is the core problem for 3D human pose tracking. In this section, we work with a model of generative pose tracking to find the conditional distribution within the framework of hierarchical evolution strategy.

5.1. Hierarchical pose tracking algorithm

To track 3D human pose, we develop an optimal tracking algorithm on the basis of $(\mu/\mu, \lambda)$-ES in conjunction with the evolutionary mechanism of AGA. The $(\mu/\mu, \lambda)$-ES [17,30] is a evolution strategy for optimization of real-valued functions $f : \mathbb{R}^N \mapsto \mathbb{R}$ that is popular both due to its proven good performance and its relative mathematical tractability. The basic idea of $(\mu/\mu, \lambda)$-ES is to select μ states from λ candidate states to create next state solution. The double appearances of the parameter μ indicate that all parents participate in the creation of every single offspring of candidate.

One of the most important issues of the $(\mu/\mu, \lambda)$-ES is how to generate the candidate states. In general, we must ensure that the true solution is in the space spanned by the candidate states. On the other hand, the space from which the candidates states are generated should be as small as possible so that we can enhance the computing efficiency. We deal with this tradeoff by means of the mechanism of AGA and the hierarchical characteristic of state space. We have introduced that the pose subspace extracted by PCA is naturally hierarchical. Given the estimated state $\mathbf{x}_t$, the next state $\mathbf{x}_{t+1}$ is reasonably in a super-ellipsoid whose

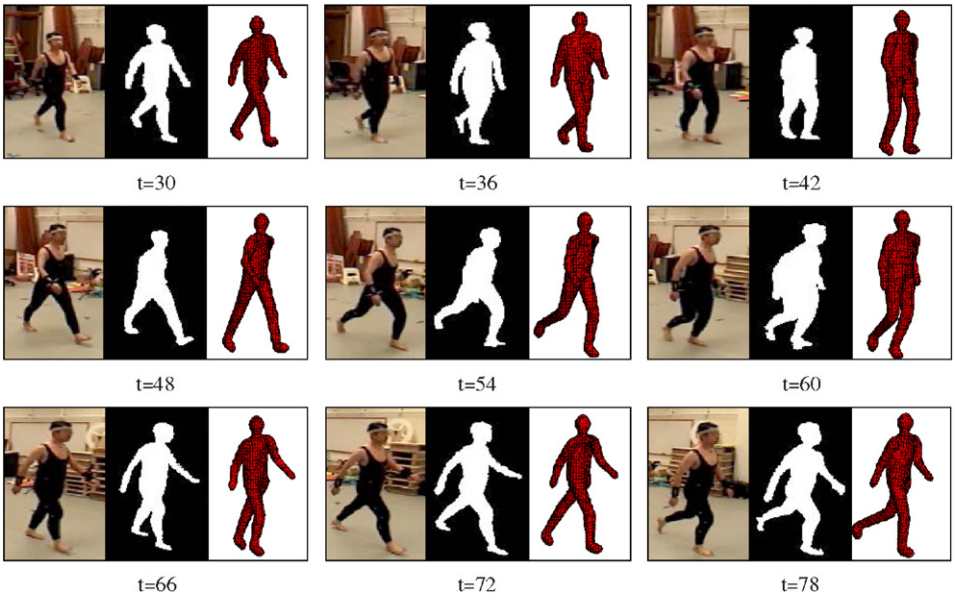

Fig. 13. Results of 3D human pose reconstruction from an image sequence in which a subject is performing a walking motion at a stride. The images are extracted from the video taken from the web site http://mocap.cs.cmu.edu/.

centroid is $\mathbf{x}_t$.[2] We can find the distribution $p(\mathbf{x}_{t+1}|\mathbf{x}_t)$ in this super-ellipsoid. To this end, we design the hierarchical pose tracking algorithm on the base of $(\mu/\mu, \lambda)$-ES and AGA. This algorithm can be considered as repeatedly updating a search point $\mathbf{x}$ using the following steps:

(1) Determine the diagonal matrix $\mathbf{C}$ according to the practical applications. Each diagonal element in $\mathbf{C}$ corresponds to a axis length of the super-ellipsoid. In our problem, $\mathbf{C}$ is relevant to the frame rate of image sequence and the topology dominance of the state components.
(2) Generate the initial mutation vector $\mathbf{z}$ consisting of d independent, standard normally distributed components, where d is the dimensionality of state space.
(3) Map $\mathbf{z}$ to problem domain: $\mathbf{x}'_t = \mathbf{x}_t + \mathbf{C}\mathbf{z}$. Determine the fitness function value $\mathscr{F}(\mathbf{x}'_t)$.
(4) Evolve the chromosome $\mathbf{z}$ according to the state evolutionary mechanism of AGA and store the μ best states $\{\mathbf{z}_i \mid i = 1, 2, \ldots, \mu\}$.
(5) Compute thearithmetic mean:

$$\langle \mathbf{z} \rangle^{(t)} = \frac{1}{\mu} \sum_{i=1}^{\mu} \mathbf{z}_i, \tag{13}$$

where, $\langle \mathbf{z} \rangle^{(t)}$ refers to progress vector.
(6) Update the search point by

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{C}\langle \mathbf{z} \rangle^{(t)}. \tag{14}$$

Among the above steps, the determination of step length diagonal matrix $\mathbf{C}$ is very important for successful pose tracking and needs to be evaluated carefully.[3]

5.2. Experiments

We demonstrate our tracking algorithm in different image sequences. As mentioned in Section 4.3, before performing the pose tracking, we extract the pose subspace from only one sequence of motion capture data for each motion type. The process of pose tracking is then executed in the subspaces. Although the using of subspace extracted from single sequence, the generative tracking framework ensures the generalization capability of our algorithm because of the effective state prediction and correction.

Generally speaking, prediction and correction are decisive steps for successful tracking. In our algorithm, there is no need to train or learn an explicit state prediction model. We generate the eligible state candidates by the usage of known hierarchical characteristic of state subspace. The super-ellipsoid topology of predictive space makes the prediction more accurate. The candidates that are agreed greatly by image observation are selected to produce next state. Another important problem of state tracking is initialization. How to begin the tracking process from a good starting point sometimes is a intractable problem. We achieve the automatic initialization by determining the pose of the first frame in the framework of HAGA, where we just view the first frame as a single image.

We test the pose tracker in image sequences describing different types of motion. The first type of motion we tested is

[2] The isotropic assumption of the search space is not suitable for our problem because the principle components of state vector $\mathbf{x}$ dominate the topology of the state space.

[3] There are some mechanisms [17] that is employed for the adaptation of the step length. Here, we do not introduce the mechanisms to our algorithm because of the known hierarchical characteristic of state space.

Fig. 14. Results of 3D human pose reconstruction from an image sequence in which a subject is performing a running motion. The images are extracted from the video taken from the web site http://mocap.cs.cmu.edu/.

Table 4
Error measures for the full body DOFs over whole sequence

		Error (in degrees)	
		Average error	Standard deviation
Walking	Sequence 1	2.8451	1.0013
	Sequence 2	2.6325	0.9710
Running	Sequence 1	2.7270	0.7095
	Sequence 2	2.9866	1.0574

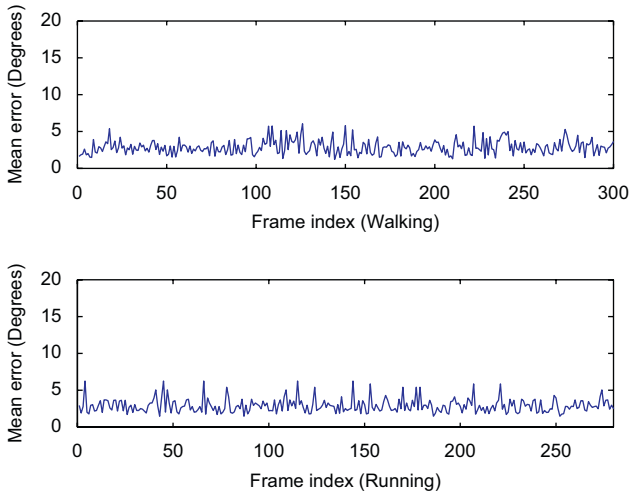

Fig. 15. The variation of mean error in test sequences.

walking. Total 316 frames of motion capture data are used to extract the motion subspace. The dimensionality of subspace is 5. The parameters of HAGA are set as $S_t = 3$, $N_t = 3$, $E_t = 5$ for careful search of the state space in initialization. To demonstrate the ability of the tracker in generalizing to different walking styles, we track the walking motion with long steps. This style is different from that of motion capture sequence from which the subspace is extracted. Fig. 13 shows the performance of pose tracking.

The second type of motion we tested is running. The frame number of motion capture data used to extract the subspace is 130. The dimensionality of subspace is 4. The parameters setup of HAGA is similar with walking motion. The tracking results can be seen in Fig. 14. Despite the coarse edges of extracted silhouette, the 3D pose tracker does a good job.

Table 4 summarizes the performance of the test sequences in walking motion and running motion. For each motion type, two sequences performed by different subjects with different frame numbers are tested. The average errors and the standard deviations over all joints angles are near 3° and 1°, respectively, in general. The mean errors over all joint angles of the test sequences are shown in Fig. 15. It can be found that the change of mean error in whole sequence is small. Our algorithm can achieve stable tracking of 3D human pose.

6. Conclusions

In this paper, we presented a novel generative approach to reconstruct 3D human pose from a single monocular image

and monocular image sequences. Our approach is a step toward describing motion characteristic of high dimensional data spaces by extracting its subspace. From motion capture data, we not only distilled the prior knowledge about human motion, but also reduced the dimensionality of problem. In the compact subspace, we perform effective search for finding the optimal poses. In static image situation, to explore the solution space efficiently, we designed the AGA and HAGA, by which the optimal solutions can be searched effectively by utilizing the characteristics of state subspace. In tracking scenario, we found the conditional state distribution $p(\mathbf{x}_{t+1}|\mathbf{x}_t)$ in the super-ellipsoid determined according to the hierarchical property of state space. We embedded the evolutionary mechanism of AGA into the framework of $(\mu/\mu, \lambda)$ evolution strategy for adapting the local characteristics of fitness function. The robust shape contexts descriptor is adopted to construct the matching function. Therefore, the validity and the robustness of the matching between image features and synthesized model features can be ensured. The approaches were tested on different human motion sequences with good results.

In terms of future work, we plan to accelerate reconstruction speed by introducing vectorize representation of human silhouette. In addition, our algorithms will be extended to cover a wilder class of human motions. The switch mechanism between different subspaces need to be explored because it is very important to deal with more complicated human motion scenario. Naturally, after coming back to 3D world from 2D images, how to recognize human motion according to the results of pose reconstruction is a considerable problem. It is also the next work we plan to investigate particularly.

Acknowledgments

This research is supported by the National Basic Research Program (973 Program) of China (No. 2006CB303103), the National Natural Science Foundation of China (No. 60675017) and the 111 project.

References

[1] C. Sminchisescu, A. Kanaujia, Z. Li, D. Metaxas, Discriminative density propagation for 3D human motion estimation, in: Proceedings of the Conference on Computer Vision and Pattern Recognition, 2005, pp. 217–323.

[2] A. Agarwal, B. Triggs, Recovering 3D human pose from monocular images, IEEE Trans. Pattern Anal. Mach. Intell. 28 (1) (2006) 44–58.

[3] R. Rosales, S. Sclaroff, Learning body pose via specialized maps, Adv. Neural Inf. Process. Syst;2002.

[4] M. Brand, Shadow puppetry, in: International Conference on Computer Vision, vol. 2, 1999, p. 1237.

[5] A. Elgammal, C. Lee, Inferring 3D body pose from silhouettes using activity manifold learning, in: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, 2004, CVPR 2004.

[6] G. Shakhnarovich, P. Viola, T. Darrell, Fast pose estimation with parameter-sensitive hashing, in: Proceedings of the Ninth IEEE International Conference on Computer Vision, 2003, pp. 750–757.

[7] J. Deutscher, A. Blake, I. Reid, Articulated body motion capture by annealed particle filtering, in: Proceedings of the 2000 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, 2000, pp. 126–133.

[8] H. Sidenbladh, M. Black, D. Fleet, Stochastic tracking of 3D human figures using 2D image motion, in: European Conference on Computer Vision, vol. 2, 2000, pp. 702–718.

[9] C. Sminchisescu, B. Triggs, Covariance scaled sampling for monocular 3D body tracking, in: IEEE International Conference on Computer Vision and Pattern Recognition, vol. 1, 2001, pp. 447–454.

[10] Y. Wu, G. Hua, T. Yu, Tracking articulated body by dynamic Markov network, in: Proceedings of the Ninth IEEE International Conference on Computer Vision, 2003, pp. 1094–1101.

[11] A. Agarwal, B. Triggs, Tracking articulated motion using a mixture of autoregressive models, in: Proceedings of the European Conference on Computer Vision, vol. 3023, 2004, pp. 54–65.

[12] H. Ning, T. Tan, L. Wang, W. Hu, People tracking based on motion model and motion constraints with automatic initialization, Pattern Recognition 37 (7) (2004) 1423–1440.

[13] G. Mori, J. Malik, Recovering 3D human body configurations using shape contexts, IEEE Trans. Pattern Anal. Mach. Intell. 28 (7) (2006) 1052–1062.

[14] H. Sidenbladh, M. Black, L. Sigal, Implicit probabilistic models of human motion for synthesis and tracking, in: European Conference on Computer Vision, vol. 1, 2002, pp. 784–800.

[15] R. Urtasun, D. Fleet, P. Fua, Monocular 3-D tracking of the golf swing, in: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2.

[16] Z. Michalewicz, Genetic Algorithms + Data Structures = Evolution Programs, Springer, London, UK, 1996.

[17] D. Arnold, H. Beyer, Optimum tracking with evolution strategies, Evol. Comput. 14 (3) (2006) 291–308.

[18] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Trans. Pattern Anal. Mach. Intell. 24 (4) (2002) 509–522.

[19] J. Aggarwal, Q. Cai, Human motion analysis: a review, Comput. Vision Image Understanding 73 (3) (1999) 428–440.

[20] D. Gavrila, Visual analysis of human movement: a survey, Comput. Vision Image Understanding 73 (1) (1999) 82–98.

[21] T. Moeslund, E. Granum, A survey of computer vision-based human motion capture, Comput. Vision Image Understanding 81 (3) (2001) 231–268.

[22] R. Urtasun, P. Fua, 3D human body tracking using deterministic temporal motion models, in: European Conference on Computer Vision, vol. 3, 2004, pp. 92–106.

[23] M. Arulampalam, S. Maskell, N. Gordon, T. Clapp, D. Sci, T. Organ, S. Adelaide, A tutorial on particle filters for online nonlinear/non-Gaussian–Bayesian tracking, IEEE Trans. Signal Proc. 50 (2) (2002) 174–188.

[24] J. Deutscher, A. Davison, I. Reid, Automatic partitioning of high dimensional search spaces associated with articulated body motion capture, in: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

[25] R. Rosales, V. Athitsos, L. Sigal, S. Sclaroff, 3D hand pose reconstruction using specialized mappings, in: IEEE International Conference on Computer Vision, 2001, pp. 378–385.

[26] C. Taylor, Reconstruction of articulated objects from point correspondences in a single uncalibrated image, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 2000.

[27] X. Jin, C. Tai, Convolution surfaces for arcs and quadratic curves with a varying kernel, Visual Comput. 18 (8) (2002) 530–546.

[28] M. Tong, Y. Liu, T.S. Huang, 3D human model and joint parameter estimation from monocular image, Pattern Recognition Lett. 28 (7) (2007) 797–805.

[29] L. Sigal, M.J. Black., Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion, Technical Report CS-06-08, Brown University.

[30] H. Beyer, The Theory of Evolution Strategies, Springer, Berlin, 2001.

About the Author—XU ZHAO is currently a Ph.D. candidate at the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University. He received the M.S. degree in electrical engineering from China Ship Research and Develop Academy, in 2004. His research interests include computer vision, machine learning, and pattern recognition. He had served as a reviewer of the international conferences ICPR and ACCV in the field of computer vision and pattern recognition.

About the Author—YUNCAI LIU received the Ph.D. degree from the University of Illinois at Urbana-Champaign, in the Department of Electrical and Computer Science Engineering, in 1990, and worked as an associate researcher at the Beckman Institute of Science and Technology from 1990 to 1991. Since 1991, he had been a system consultant and then a chief consultant of research in Sumitomo Electric Industries Ltd., Japan. In October 2000, he joined the Shanghai Jiao Tong University as a distinguished professor. His research interests are in image processing and computer vision, especially in motion estimation, feature detection and matching, and image registration. He also made many progresses in the research of intelligent transportation systems.