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.