Pattern Recognition Letters 30 (2009) 1077–1085

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Action recognition feedback-based framework for human pose reconstruction from monocular images

Xinxiao Wu, Wei Liang *, Yunde Jia

Beijing Laboratory of Intelligent Information Technology, School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, PR China

ARTICLE INFO

Article history:
Available online 9 April 2009

Keywords:
Human pose reconstruction
Action recognition feedback
Motion correlation
Manifold motion template

ABSTRACT

A novel framework based on action recognition feedback for pose reconstruction of articulated human body from monocular images is proposed in this paper. The intrinsic ambiguity caused by perspective projection makes it difficult to accurately recover articulated poses from monocular images. To alleviate such ambiguity, we exploit the high-level motion knowledge as action recognition feedback to discard those implausible estimates and generate more accurate pose candidates using large number of motion constraints during natural human movement. The motion knowledge is represented by both local and global motion constraints. The local spatial constraint captures motion correlation between body parts by multiple relevance vector machines while the global temporal constraint preserves temporal coherence between time-ordered poses via a manifold motion template. Experiments on the CMU Mocap database demonstrate that our method performs better on estimation accuracy than other methods without action recognition feedback.

1. Introduction

Inferring 3D poses of a complex articulated human body from monocular images has received significant attention in computer vision. It is of broad interest for a very wide range of applications, e.g. visual surveillance, clinical studies and human computer interactions. This paper addresses specifically the problem of reconstructing articulated poses from images using a monocular camera rather than a stereo or multi-camera system.

Recovering 3D articulated poses from monocular 2D images is geometrically under-defined. Projection a 3D scene onto a 2D image plane suppresses depth information and thus makes the solution of pose reconstruction ambiguous. Many previous methods introduced temporal continuity (Agarwal and Triggs, 2004a,b; Moon and Pavlovic, 2006; Wang et al., 2008), multiple hypotheses (Agarwal and Triggs, 2005; Sminchisescu et al., 2005; Thayananthan et al., 2006), and the combination of top-down and bottom-up processing (Rosales and Sclaroff, 2006; Sminchisescu et al., 2006) to handle the ambiguity problem. An important source of information, which has not been exploited, is the high-level motion knowledge of action recognition feedback, and we attempt to incorporate such motion information to rule out those implausible pose hypotheses and yield more accurate pose candidates from prior distributions.

To represent and learn the motion knowledge, two types of constraints in natural human movement are exploited. One is the global temporal constraint, that is, the occurrence of poses within a motion should follow some specific order (i.e. temporal context). Taking "picking up the ball" for example, reaching a hand to grasp the ball should occur between bending down and standing up straight. Another is the local spatial constraint, that is, the body parts should have dependence on each other (i.e. motion correlation). For example when "walking", the right leg steps forward while the right arm swings backward and left-arm swings forward. With the global constraint, we model each action class as a trajectory on the low-dimensional manifold space consisting of sequential poses, called the manifold motion template, to represent the global motion information. This motion template provides the prior distribution over 3D poses on manifold space. With the local constraint, we utilize multiple Relevance Vector Machines (RVMs) to establish non-linear relationships between body parts of each action class to define the local motion knowledge.

1.1. Action recognition feedback-based framework

The proposed framework based on action recognition feedback includes three modules: 2D-to-3D, 3D-to-semantic and semantic-to-3D. In the 2D-to-3D module, 3D articulated poses of input motion are roughly recovered from 2D images. In the 3D-to-semantic module, the motion is recognized from the sequence of poses recovered in the 2D-to-3D module. In the semantic-to-3D module,

* Corresponding author. Tel./fax: +86 10 6891 4849.
E-mail addresses: wuxinxiao@bit.edu.cn (X. Wu), liangwei@bit.edu.cn (W. Liang), jiayunde@bit.edu.cn (Y. Jia).

0167-8655/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2009.04.002

the action recognition feedback is incorporated to correct the previous estimates.

In this paper, we mainly focus on the modules of 3D-to-semantic and semantic-to-3D. In the 2D-to-3D module, a Bayesian Mixture of Experts (BME) (Sminchisescu et al., 2005) is adopted on the input image space to calculate several possible poses with their corresponding probabilities for each frame. For notation clarity, the poses recovered in the 2D-to-3D module are called *rough poses* in the rest paper. In the 3D-to-semantic module, the maximum a posteriori (MAP) method is employed to recognize the motion through calculating the modified Huasdorff distances between the sequence of rough poses and motion templates of all the classes. In the semantic-to-3D module, we incorporate both the local RVMs motion correlation and the global manifold motion template as action recognition feedback to update the rough poses. The rough poses are firstly refined by RVMs and then combined with new pose candidates generated from the motion template according to the image observation.

Fig. 1 illustrates a schematic graph of the action recognition feedback-based framework for articulated pose recovery which consists of three main processes: mapping from 2D images to 3D rough poses, recognizing the motion from 3D rough poses and feeding back both local and global motion knowledge. In Fig. 1, there are two loop processes: 2D-to-3D-to-2D and 3D-to-semantic-to-3D. Many methods for recovering articulated pose work in the 2D-to-3D-to-2D loop. For example, the data-driven discriminative methods (Agarwal and Triggs, 2006; Sminchisescu et al., 2005) correspond to the mapping from 2D image to 3D pose; The model-driven generative methods (Wang et al., 2008; Hou et al., 2007) correspond to the projection of possible poses to the image plane for searching the optimal pose; the combination of top-down and bottom-up methods (Rosales and Sclaroff, 2006; Sminchisescu et al., 2006) correspond to the whole loop of discriminative and generative processes. Different from these methods, our method introduces the 3D-to-semantic-to-3D loop which utilizes the feedback knowledge of action recognition for pose estimation to alleviate the ambiguity problem.

In order to evaluate the proposed method, we conduct experiment on the golf motions of CMU Mocap database where both images and 3D poses are simultaneously captured. The 3D pose state is represented by 59 joint angles and we choose 49 joint angles which exhibit large movement to be the complete pose vector. Background images are clean with low noise and background subtraction based on a mixture of Gaussians is utilized to extract the human silhouettes for image features.

The remainder of this paper is organized as follows. The related work is presented in Section 2. Section 3 and Section 4 respectively describe the local spatial motion correlation and the global manifold motion template. The pose estimation algorithm based on action recognition feedback is given in Section 5. The experimental results are demonstrated in Section 6. We draw conclusions and discuss future work in Section 7.

2. Related work

Many existing approaches are proposed to handle the ambiguity problem in 3D pose reconstruction from 2D images. Some dynamic models are introduced to exploit temporal continuity hidden in motion because pose recovery is often addressed in the context of a sequence of images. Agarwal and Triggs (2004a) fused an autoregressive dynamic model with the image observations for estimating the body pose in a regression-based setting. They also proposed a mixture of Gaussian autoregressive process to model the non-linear and time-varying dynamics (Agarwal and Triggs, 2004b). Wang et al. (2008) exploited prior motion knowledge about typical body poses using a Gaussian Process Dynamical Model (GPDM) for monocular 3D people tracking. Many of these prior motion models are derived from training data accounting for variations in movements. Nevertheless, natural human movement is very complex with different actions performing different dynamics. Also, most current motion models are limited in some specific movements such as walking or jogging.

An alternative to handle the ambiguity is to maintain multiple hypotheses of pose solutions for a given image because the mapping from 2D image to 3D pose is multi-valued. A mixture of regressors (Agarwal and Triggs, 2005) has been proposed to explicitly calculate several possible poses from monocular image under the assumption that each regressor spans a small set of examples to reduce ambiguities. Thayananthan et al. (2006) learn a set of relevance vector machines mapping functions to estimate the distribution of possible configurations from a single frame. Sminchisescu et al. (2005) discuss a mixture density propagation algorithm with a Bayesian mixture of experts to estimate 3D human motion in monocular video sequences. These multiple hypothesis methods could handle a one-to-many mapping, but they depend only on the bottom image information. It results in the failure of recovery when the image feature (body silhouette) is not clearly extracted with noise.

Another solution to solve the ambiguity problem is by integrating top-down with bottom-up processing. Rosales and Sclaroff (2006) propose a probabilistic, non-linear supervised computation learning model: Specialized Mapping Architecture (SMA), combining a forward model and a feedback model. The forward process infers several possible poses from an input image and a feedback model projects the possible poses back to the image plane for observation to discard implausible poses to select more accurate poses. Sminchisescu et al. (2006) present an algorithm for jointly learning a consistent bidirectional generative-recognition model for monocular pose reconstruction. A simple but effective feedback

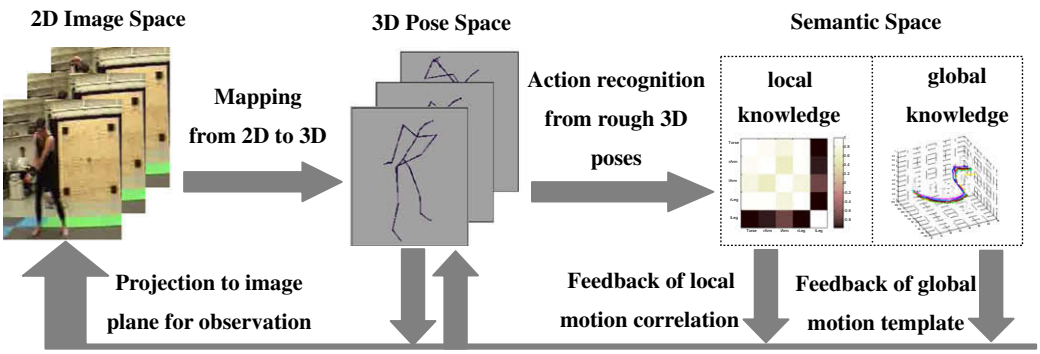

Fig. 1. The framework based on action recognition feedback for articulated human pose reconstruction.

gate based on an approximate generative observation model is introduced to reduce incorrect estimates inferred only from a discriminative model. This feedback is still simple and does not exploit extra prior motion knowledge to guide the pose recovery.

3. Local spatial motion correlation of action recognition feedback

During natural human motion, each body part is not independent, but dependent on other body parts. Hence we employ multiple Relevance Vector Machines (RVMs) regression to preserve non-linear relationships between body parts as the local motion knowledge.

The complete body represented by 49-dimensional joint angle vector is partitioned into five parts, including torso, two arms and two legs. Accordingly, the 49-dimension vector is divided into five sub-vectors of which each represents a body part (e.g. the first to the 21st-dimension of the 49-dimensional joint angle vector denotes the torso part, the 22nd to the 28th-dimension denotes the left-arm part). Thus the multiple RVMs model actually establishes the non-linear mapping between joint angle vectors. Nevertheless, not all the pairs of body parts have strong correlations of a specific action class, i.e. some two-part correlations are strong and some are weak. Therefore, we just choose strong two-part correlations to describe the local motion knowledge. In "full swing", for example, the left leg has strong correlation with the right leg but less dependence on the two arms, so we select the pair of left leg and right leg, neglecting the pair of left leg and left-arm as well as the pair of left leg and right arm.

3.1. Multiple RVMs for local motion correlation

We arrange T sequential observations of two body parts into two matrices, $P = [p_1, p_2, \cdots p_T] \in R^{D_p \times T}$ and $Q = [q_1, q_2, \cdots q_T] \in R^{D_q \times T}$, where the column vectors $p_t \in R^{D_p \times 1}$ and $q_t \in R^{D_q \times 1}$ respectively correspond to joint angle vectors of two body parts P and Q at time t, D_p and D_q respectively denote the state-space dimensions of body parts P and Q. The motion correlation coefficient between P and Q is defined by

$$correlation(P, Q) = \frac{|cov(P', Q')|}{\sqrt{var(P')}\sqrt{var(Q')}}, \tag{1}$$

where $P' = TP$ and $Q' = UQ$, with $T \in R^{1 \times D_p}$ and $U \in R^{1 \times D_q}$ respectively indicating the dimension reduction vectors to obtain one-dimensional subspace of p_t and q_t. We utilize PCA to learn the subspaces so T and U are actually the eigenvectors. Fig. 2 demonstrates the correlation maps of five body parts in two examples of "full swing" and "pick up". From the correlation maps, it is obvious that different actions have different motion correlations between body parts and the same pair of body parts maintains different correlations in different actions. The pairs of body parts with high correlation coefficients are selected to represent the local semantic knowledge according to the correlation map.

If two body parts with strong motion correlation have been selected, we apply multiple Relevance Vector Machines (RVMs) (Tipping, 2001) to learn the non-linear mapping between them. This differs from (Xu and Li, 2007) which uses linear partial least square regression method in the case of simple walking and jogging. RVM is a Bayesian regression framework and has good generalization performance on regression. A set of RVMs is adequate to capture the complex non-linear mapping from one body part to the other. Different action classes have different multiple RVMs to model the local correlations between body parts.

To learn a multiple RVMs model between body parts p and q, we collect training data consisting of N pairs of joint angle vectors $\{p_n, q_n\}_{n=1}^N$ from different motion instances of the same class varying in the execution time and motion style, where $p_n \in R^{D_p \times 1}$ and $q_n \in R^{D_q \times 1}$ respectively represent the joint angle vectors of body parts p and q. Then the regression by learning K different functions can be formulated as

$$p_n = \sum_{k=1}^{K} W^k f(q_n) + \xi^k, \tag{2}$$

where $f(q_n) = [\phi_1(q_n), \phi_2(q_n), \ldots, \phi_s(q_n)]^T$ is a vector of basis functions with S indicating the number of basis functions. The weights of the basis functions are written in matrix form. ξ^k is a Gaussian noise vector with 0 mean and diagonal covariance matrix ψ^k. GMM with K modes is applied to partition the whole training set into K subsets. Given the matrix C where the element $c_{k,n}$ is the probability that sample point q_n belongs to mapping function k, we develop a maximum likelihood estimation to efficiently compute the parameters $\{W^k, \psi^k\}$ through minimizing the following cost function:

$$L = \sum_{k=1}^{K} \sum_{n=1}^{N} c_{k,n} (p_n - W^k f(q_n))^T \psi^k (p_n - W^k f(q_n)). \tag{3}$$

Letting the first derivative of L with respect to W^k be zero, W^k can be calculated in closed form as

$$W^k = \left(\sum_{n=1}^{N} c_{k,n} p_n f(q_n)^T \right) \left(\sum_{n=1}^{N} c_{k,n} f(q_n) f(q_n)^T \right)^{-1}. \tag{4}$$

Then ψ^k is given by

$$\psi^k = diag\left(\sum_{n=1}^{N} c_{k,n} (p_n - W^k f(q_n))^T (p_n - W^k f(q_n)) \Big/ \sum_{n=1}^{N} c_{k,n} \right). \tag{5}$$

Fig. 3 illustrates an example of motion correlation model between right leg and left leg in "pick up". The motion of the right leg is accurately predicted from the left leg through the learned multiple RVMs.

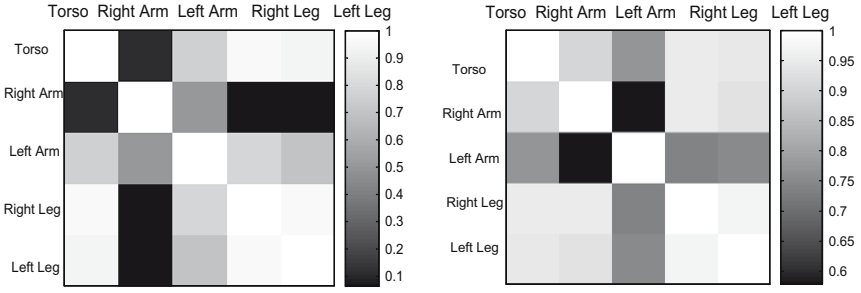

Fig. 2. Two examples of the correlation map between different body parts: full swing (left) and pick up (right). A light grid indicates strong motion correlation and a dark grid represents week correlation.

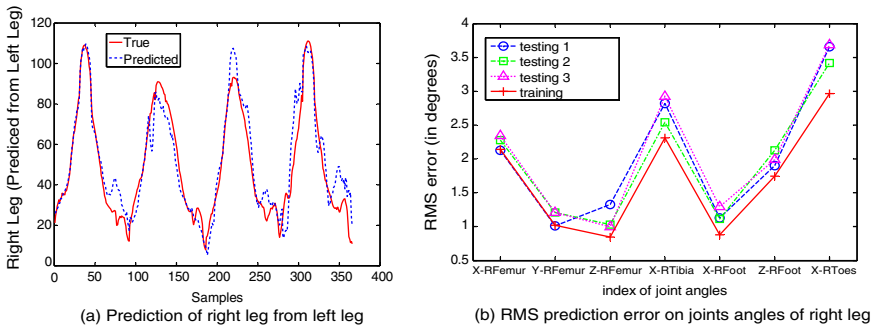

Fig. 3. The multiple RVMs correlation model between right leg and left leg. (a) shows the evolution of predicted right leg (plotted by dashed line) from left leg and the corresponding ground truth of right leg (plot by solid line) in testing sequences; The horizontal axis indexes the testing frames and the vertical axis represents the latent coordinates of the right leg; (b) illustrates RMS (Agarwal and Triggs, 2006) prediction error on joint angles of right leg compared with ground truth, where the horizontal axis indexes the joint angles and the vertical axis represents the RMS error in degrees.

3.2. Integrating local spatial motion correlation

Suppose that 3D rough poses at time t are represented by $\{x_t^i\}_{i=1}^{N_e}$, where $x_t^i \in R^{D\times 1}$ is the joint angle vector of complete body pose ($D = 49$). Given the motion correlations between body parts C^{local}, the probability of x_t^i after integrating motion correlations is given by

$$p(x_t^i|C^{local}) \propto \sum_p \sum_q p(x_{t,p}^i|x_{t,q}^i,\theta_{pq})T_{pq} = \sum_p \sum_q \left(\sum_k N(x_{t,p}^i|W_{pq}^k f(x_{t,q}^i),\xi^k)\right)T_{pq}, \quad i = 1,2,\ldots,N_e, \tag{6}$$

where $x_{t,p}^i \in R^{D_p\times 1}$ and $x_{t,q}^i \in R^{D_q\times 1}$ respectively indicate the joint angle vectors of body parts p and q in the complete body x_t^i. $\theta_{pq} = \{W_{pq}^k, \xi_{pq}^k\}_{k=1}^K$ denotes the parameters of RVMs between p and q. T_{pq} is a binary variable indicating whether the correlation between p and q is chosen or not, defined by

$$T_{pq} = \begin{cases} 1 & \text{pair of } p \text{ and } q \text{ is chosen} \quad correlation(x_{t,p}^i, x_{t,q}^i) \geqslant \varepsilon, \\ 0 & \text{pair of } p \text{ and } q \text{ is not chosen} \quad correlation(x_{t,p}^i, x_{t,q}^i) < \varepsilon. \end{cases} \tag{7}$$

Obviously, the more similar a pose's motion dependence between body parts to the feedback correlation model, the higher its probability will become.

4. Global temporal motion template of action recognition feedback

Global temporal motion templates provide temporal relationships between sequential complete poses. Due to the high-dimension of the complete pose vector, motion is represented by a pose trajectory on the low-dimensional manifold space. Although pose sequences of the same class are similar, no sequence is exactly the same for the noise and variation among different instances. Consequently, we include variance in the motion template by allowing deviations from the mean motion, and establish a Gaussian probability distribution over the low-dimensional poses at each frame. Thus the motion template can be viewed as a chain of Gaussians over time.

During the feedback of global motion knowledge, new pose candidates are generated from the motion template (i.e. sampled from the Gaussian distribution over poses at each frame) and then combined with previous rough poses according to their mixing weights. The mixing weights are derived by projecting poses back to the image plane using generalized radial basis function (GRBF) for feature measuring.

4.1. Manifold motion template for global temporal relationship

Neighbor preserving embedding (NPE) (He et al., 2005) is proposed as a non-linear dimension reduction method to learn the low-dimensional representation of high-dimensional data by capturing the intrinsic structure of data, and we adopt NPE to learn a compact description of high-dimensional 3D motion, i.e. the trajectory of poses on the low-dimensional space. Because the motion trajectories of the same action have different sequence lengths, we use Spline interpolation to align the trajectories to the same length and take the mean trajectory with variance as the manifold motion template. The dimension of low-dimensional manifold space is set to one, which is enough to efficiently express the pose information. More detailed learning of the motion template of each action class is listed in Algorithm 1. Fig. 4 shows two examples of temporal manifold motion templates.

Algorithm 1. Learning the temporal manifold motion template

Input: M motion instances $X_1, X_2, \ldots, X_M$ of the same class, where $X_i = [x_{i1}, x_{i2}, \ldots, x_{iT_i}] \in R^{D\times T_i}$ is a sequence of complete body poses, $x_{it} \in R^{D\times 1}$ ($D = 49$) represents the joint angle vector at time t of instance i, and T_i is the length of motion instance X_i.

Output: The motion template $Y = [y_1, y_2, \ldots, y_T]$, where $y_t \sim N(\mu_t, \sigma_t^2)$ and T is the length of the template

1. Learn the manifold space. Utilize NPE to learn the one-dimensional manifold representation Y_i of X_i by $X_i \to Y_i$: $Y_i = AX_i$, where $A \in R^{1\times D}$ is the dimension reduction matrix and $Y_i = [y_{i1}, y_{i2}, \ldots, y_{iT_i}] \in R^{1\times T_i}$.
2. Compute the motion template. Normalize Y_i to the same length T by Spine interpolation. Then calculate the mean trajectory $\{\mu_1, \mu_2, \ldots, \mu_T\}$ by $\mu_t = \sum_{i=1}^M y_{it}/M$, and the variance $\{\sigma_1, \sigma_2, \ldots, \sigma_T\}$ by $\sigma_t = \sum_{i=1}^M (y_{it} - \mu_t)^2/M$.

4.2. Integrating global temporal motion template

The rough poses and the pose candidates from the global motion template are projected to the image plane for observation and to obtain their mixing weights. The generalized radial basis function (GRBF) interpolation (Elgammal and Lee, 2004) is used as a projection function from low-dimensional pose space to image (silhouette) space.

Let $\{y_i \in R^{d\times 1}, i = 1,2,\ldots,N\}$ be poses represented on low-dimensional pose space and the corresponding silhouettes be $\{z_i \in R^{m\times 1}, i = 1,2,\ldots,N\}$, where d is the dimension of the pose space ($d = 1$) and m is the dimension of the silhouette space. Let $T = \{t_j \in R^{d\times 1}, j = 1,\ldots,N_t\}$ be a set of N_t centers in the manifold pose space, which are obtained by k-means clustering. The non-linear projection

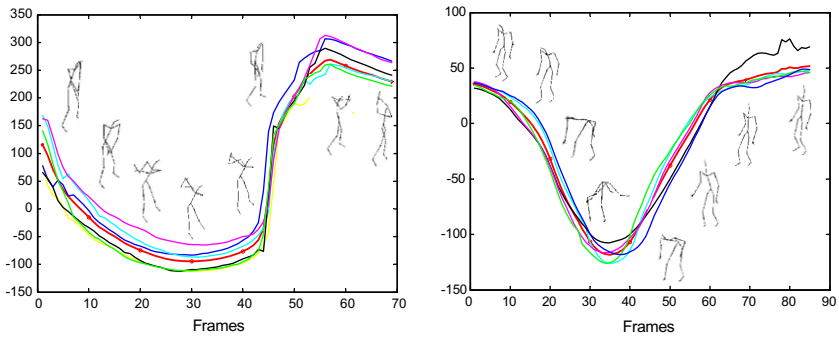

Fig. 4. Two examples of temporal motion templates represented on low-dimensional manifold space: full swing (left) and picking up (right). The horizontal axis indicates the frame numbers and the vertical axis indicates the one-dimension manifold coordinates of 3D poses. Different dashed lines correspond to different training instances and the solid line represents the mean trajectory. The prior distribution of the pose at each frame is represented by a Gaussian model, and consequently the motion template is established by a chain of Gaussian distributions.

function from the low-dimensional pose space to the kth-dimension (pixel) in the silhouette space is learned by

$$z_i^k = f^k(y_i) = p^k(y_i) + \sum_{j=1}^{N_t} v_j^k \phi(|y_i - t_j|), \tag{8}$$

where $\phi(\cdot)$ is a real-valued basis function, v_i^k are real coefficients and $p^k(y_i) = [1 y_i^T] \cdot c^k$ is a linear polynomial with coefficients c^k. The whole mapping can be written as

$$z_i = B \cdot \varphi(y_i), \tag{9}$$

where B is a $m \times (N_t + d + 1)$ matrix with the k-th row $[v_1^k \cdots v_{N_t}^k c^{kT}]$ and the vector $\varphi(y_i)$ is $[\phi(|y_i - t_1|) \cdots \phi(|y_i - t_{N_t}|) 1 y_i^T]^T$. By imposing the additional constraints $\sum_{j=1}^{N_t} v_j^k [1 t_j^T]^T = 0$, the solution for B can be obtained by directly solving the linear system

$$\begin{pmatrix} E & P_x \\ P_t^T & 0 \end{pmatrix} B^T = \begin{pmatrix} Z \\ 0 \end{pmatrix}, \tag{10}$$

where E is $N \times N_t$ matrix with $E_{ij} = \phi(|y_i - t_j|)$, P_x is a $N \times (d+1)$ matrix with ith row $[1 y_i^T]$, P_t is a $N_t \times (d+1)$ matrix with jth row $[1 t_j^T]$, and $Z = [z_1 \ldots z_N]^T$ represents the input silhouettes.

At time t, given the GRBF projection function, the input silhouette image $z_t \in R^{m \times 1}$, the global pose candidates $y_t^s \in R$ sampled from $N(u_t, \sigma_t^2)$ and the rough pose $y_t^e \in R$ represented on the one-dimensional pose space, the mixing weight w_t^s of y_t^s and the mixing weight w_t^e of y_t^e can be calculated by

$$\begin{aligned} w_t^s &\propto \exp\{-|z_t - B \cdot \varphi(y_t^s)|^2/(2\sigma^2)\}, \\ w_t^e &\propto \exp\{-|z_t - B \cdot \varphi(y_t^e)|^2/(2\sigma^2)\}. \end{aligned} \tag{11}$$

5. Articulated pose reconstruction based on action recognition feedback

We have proposed how to represent and incorporate the local and global motion knowledge of action recognition feedback. Nevertheless, the feedback should be based on the action recognition result, in other words, we should firstly get the action class label of the testing motion from its rough pose sequence and then incorporate the motion knowledge of the corresponding action class.

5.1. MAP algorithm for action recognition

As the output of 2D-to-3D module, the testing motion is represented by several rough 3D poses at each frame. We compute the mean pose sequence of the testing motion (i.e. the mean pose at each frame is the expectation of the rough poses), and project it to the manifold pose space of each action class to obtain the testing trajectory. The motion is recognized based on MAP by calculating the modified Hausdorff distances (Huttenlocher et al., 1993) between testing trajectories and motion templates. The procedure of action recognition is shown in Algorithm 2.

Algorithm 2. Action recognition by MAP

Input: The mean pose sequence $\overline{X} = [\overline{x_1}, \overline{x_2}, \ldots, \overline{x_T}] \in R^{D \times T}$ of the testing motion, where $\overline{x_t} \in R^{D \times 1} (D = 49)$ represents the mean complete pose vector at time t, T is the sequence length; Maction classes represented by $\{C_1, \ldots, C_M\}$ with $C_m = (A_m, U_m)$, where $A_m \in R^{1 \times D}$ (calculated by NPE) denotes the dimension reduction matrix for one-dimensional manifold pose space of class m and $U_m = [\mu_{m,1}, \mu_{m,2}, \ldots, \mu_{m,T}] \in R^{1 \times T}$ (output of Algorithm 1) indicates the mean pose trajectory in the motion template of class m

Output: The class label of $\overline{X}$

For $m = 1$ to M Do

1. Calculate the one-dimensional trajectory of $\overline{X}$ on the manifold space of class$m$: $Y_m = [y_{m,1}, y_{m,2}, \ldots, y_{m,T}] \in R^{1 \times T}$ with $y_{m,t} = A_m \overline{x_t}$.
2. Compute the modified Hausdorff distance between Y_m and U_m by: $Hausdorff(Y_m, U_m) = \sum_{t=1}^{T} \max(H_{Y_m}(t), H_{U_m}(t))$, where $H_{Y_m}(t) = \min_{-\varepsilon \leqslant r \leqslant \varepsilon}(\|y_{m,t} - \mu_{m,t+r}\|)$ and $H_{U_m}(t) = \min_{-\varepsilon \leqslant r \leqslant \varepsilon}(\|\mu_{m,t} - y_{m,t+r}\|)$. The class-conditional probability is given by $p(\overline{X}|C_m) \propto \exp(-Hausdorff(Y_m, U_m)/\|Y_m\|)$.
3. The class label j is calculated by maximizing the posterior probability

$$j = \arg \max_{m=1,2,\ldots,M} p(C_m|\overline{X}), \quad p(C_m|\overline{X}) \propto p(\overline{X}|C_m) p(C_m).$$

Fig. 5 shows four one-dimensional manifold pose spaces of four action classes (a) full swing, (b) pick up, (c) place ball, and (d) place tee. The horizontal axis indexes the frames and the vertical axis represents the coordinates of manifold space. In each figure, the solid dotted line represents the motion template of the corresponding class (note that we just illustrate the mean pose trajectory of the motion template without pose variance). For example, the solid dotted line in Fig. 5a depicts the motion template of full swing. There are four testing motions: "full swing", "pick up", "place ball" and "place tee", to be recognized. Taking the testing "pick up" for example: it has four projected testing trajectories in four manifold spaces, plotted by dash dotted lines in (a)–(d), and the most matched motion template with the nearest Hausdorff distance is pick up in (b), so the class label is pick up.

5.2. Combining local and global motion knowledge feedback

After the testing sequence is recognized, we apply local spatial motion correlation to update the probabilities of rough poses, and combine these updated poses with new pose candidates generated from the global temporal motion template according to their mixing weights to obtain the final pose estimation. The detailed procedure of combining local and global motion knowledge feedback is listed in Algorithm 3.

Algorithm 3. Combining local and global motion knowledge feedback

Input: A sequence of roughly recovered 3D poses (the output of the 2D-to-3D module) $\{\{x_1^j, w_1^j\}_{j=1}^{N_e}, \{x_2^j, w_2^j\}_{j=1}^{N_e}, \ldots, \{x_T^j, w_T^j\}_{j=1}^{N_e}\}$, where $x_t^j \in R^{D\times 1}(D = 49)$ represents the complete 3D pose vector, w_t^j denotes the corresponding probability and N_e the number of rough poses for each frame; A sequence of images $Z = \{z_1, z_2, \ldots, z_T\}$, where $z_t \in R^{m\times 1}$ represents the silhouette at time t; the local motion correlation parameters $\{W_{pq}^k, \xi_{pq}^k, T_{pq}\}_{k=1}^K$; the global manifold motion template $Y = \{y_1, y_2, \ldots, y_T\}$ with $y_t \sim N(\mu_t, \sigma_t^2)$ and the GRBF projection function parameter B;

Output: 3D pose sequence$\{x_1, x_2, \ldots, x_T\}$ with $x_t \in R^{D\times 1}$

For $t = 1$ to T do

1. Feed back the local motion correlation. Utilize the local correlation modeled by multiple RVMs to update w_t^i.

   For $i = 1$ to N_e do

   $$w_t^i \propto w_t^i \times \sum_p \sum_q \left(\sum_k N(x_{t,p}^i | W_{pq}^k f(x_{t,q}^i), \xi_{pq}^k) \right) T_{pq}, \quad \sum_i w_t^i = 1.$$

   End

2. Feed back the global motion template. Generate N_s pose samples $\{y_t^j, v_t^j\}_{j=1}^{N_s}$ from $N(u_t, \sigma_t^2)$, where $y_t^j \in R$ is the pose represented in one-dimensional manifold space with its probability v_t^j. Rough poses x_t^i are projected to the manifold space to find the corresponding manifold representation $s_t^i \in R$. The probabilities of y_t^j and s_t^i are updated by image observation.

   For $j = 1$ to N_s do

   $$v_t^j \propto v_t^j \times \exp\{-|z_t - B \cdot \varphi(y_t^j)|^2/(2\sigma^2)\}, \quad \sum_j v_t^j = 1.$$

   End

   For $i = 1$ to N_e do

   $$w_t^i \propto w_t^i \times \exp\{-|z_t - B \cdot \varphi(s_t^i)|^2/(2\sigma^2)\}, \quad \sum_i w_t^i = 1.$$

   End

3. Compute the final reconstructed pose by

   $$x_t = F\left(\left(\sum_{i=1}^{N_e} s_t^i \times w_t^i + \sum_{j=1}^{N_s} y_t^j \times v_t^j\right) / \left(\sum_{i=1}^{N_e} w_t^i + \sum_{j=1}^{N_s} v_t^j\right)\right),$$

   where F is an inverse function from the one-dimensional manifold space to the original D-dimensional ($D = 49$) complete pose space.

End

6. Experiments

In this section, we describe the 3D articulated pose reconstruction in real test sequences and give comparisons with existing methods. For all experiments, the non-optimized Matlab code is running on a Dell PC with Intel Pentium D 3.4 GHz CPU and 1 GB RAM.

6.1. Pose representation and image descriptor

The experiments are conducted on the CMU MoCap Database (see the website at: http://mocap.cs.cmu.edu/), which simultaneously captures 2D visual images and 3D pose configurations represented by 59 joint angles of an articulated human body. We select 49 joint angles that exhibit large movement to represent the complete body pose. The background subtraction using a Gaussian mixture model (Stauffer and Grimson, 1999) is applied to extract the foreground human regions, and then the silhouette is obtained by removing noise and filling in regions. For computational efficiency, the silhouette image is normalized to a size of 75×90pixels. Fig. 6a shows an input image, Fig. 6b shows the extracted foreground human region, and Fig. 6c the normalized silhouette image.

Four golf actions of around 1200 frames are chosen, including full swing, picking up, placing ball and placing tee, each performed by several times. The proposed method is capable of dealing with more than four classes, where the additional motion knowledge of the increasing classes should be learned offline in training phase. For each action class, we select half of the data for training the global motion template as well as local motion correlation, and the remainder for testing. Both training and testing motion sequences are manually segmented prior to processing and the action recognition is based on the whole motion sequence.

6.2. Implementation and construction results

The offline training procedure includes the BME for mapping from 2D to 3D, multiple RVMs for local motion correlation, manifold motion templates for global temporal relationship, and the GRBF projection function for image observation. The online testing process consists of four steps: firstly rough 3D poses of the testing motion are recovered by BME, secondly the motion is recognized from sequence of rough poses by MAP, thirdly multiple RVMs are utilized to update the rough poses by recalculating their probabilities, and finally the refined poses after RVMs as well as the candidate poses sampled from motion template are projected to the image plane by GRBF for observation to update their mixing weights.

In BME, five mixture experts are learned to calculate the five rough 3D complete poses at each frame (i.e.$N_e = 5$ in Algorithms 2 and 3). In RVMs, we simply use the linear basis functions (i.e. $\phi_i(q) = i$th component of q), and the number of RVMs varies from 6 to 30 in different mappings. During the global semantic feedback, five pose candidates are generated at each frame (i.e. $N_s = 5$) from the manifold motion template and the biharmonic basis function (i.e. $\phi(|y_i - t_j|) = |y_i - t_j|$) is used in GRBF. In terms of computational time, the online 3D pose recovery costs 0.047 second per frame mainly depending on the number of rough poses (i.e. N_e) and pose candidates from the global motion template (i.e. N_s). Silhouette extraction is currently done offline, but would be possible online in real-time.

Fig. 7 intuitively demonstrates the reconstruction results of 3D pose sequences in real golf motion. We adopt the animation tool Maya (see the website at: http://www.alias.com/) to illustrate the skeleton representation of recovered 3D articulated pose. There are two sets of images, for each the top row shows the original input images and the bottom row is the corresponding reconstruction results represented by a sketch model. Although some human silhouettes are not clearly extracted because the foreground has similar color to the background, our method is still able to recover the articulated human pose with good accuracy even in the case of self-occlusion.

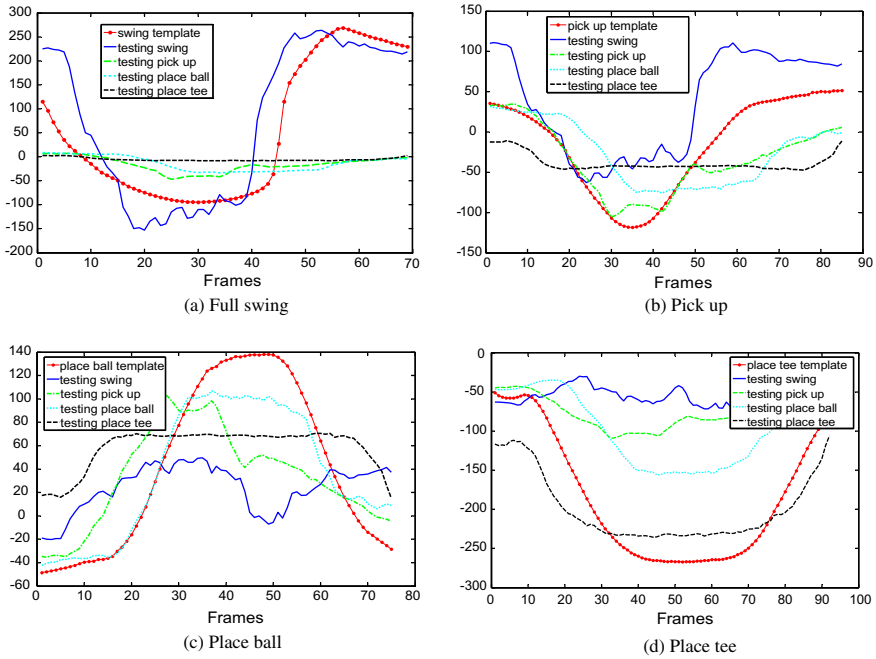

Fig. 5. Action recognition based on the manifold motion template. The horizontal axis is the frame number and the vertical axis is the one-dimensional manifold coordinate of 3D poses. The solid dotted line indicates the manifold motion template (without showing the variance).

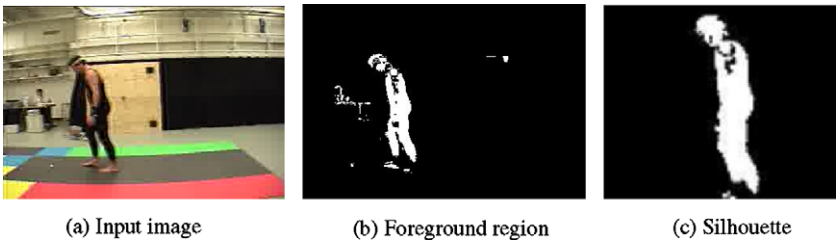

Fig. 6. Silhouette image extraction.

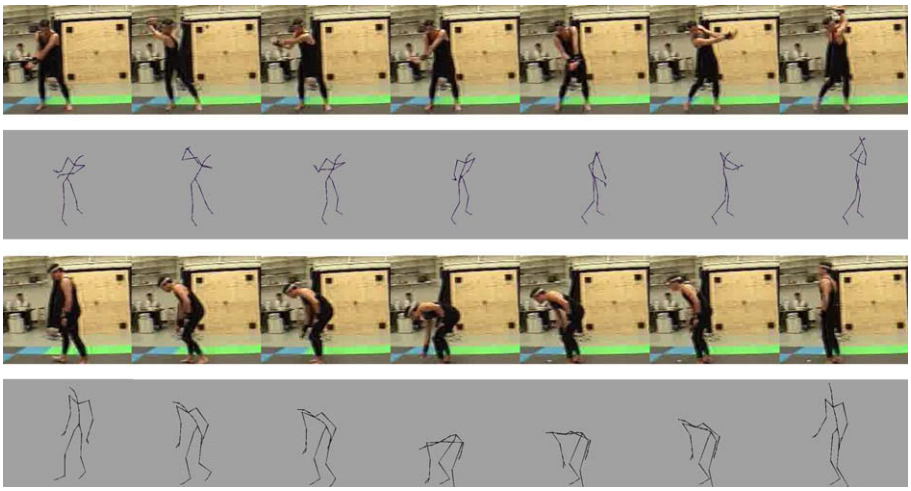

Fig. 7. Reconstructions of 3D human pose by Maya.

6.3. Comparative evaluation

We compare our method with the mixture of regressors (Agarwal and Triggs, 2005) and combination of top-down with bottom-up method (Rosales and Sclaroff, 2006). Fig. 8 compares the mean RMS error of all the joint angles over four testing motions and Table 1 concludes the comparisons on average and maximum errors in degrees. Different from other methods, the proposed

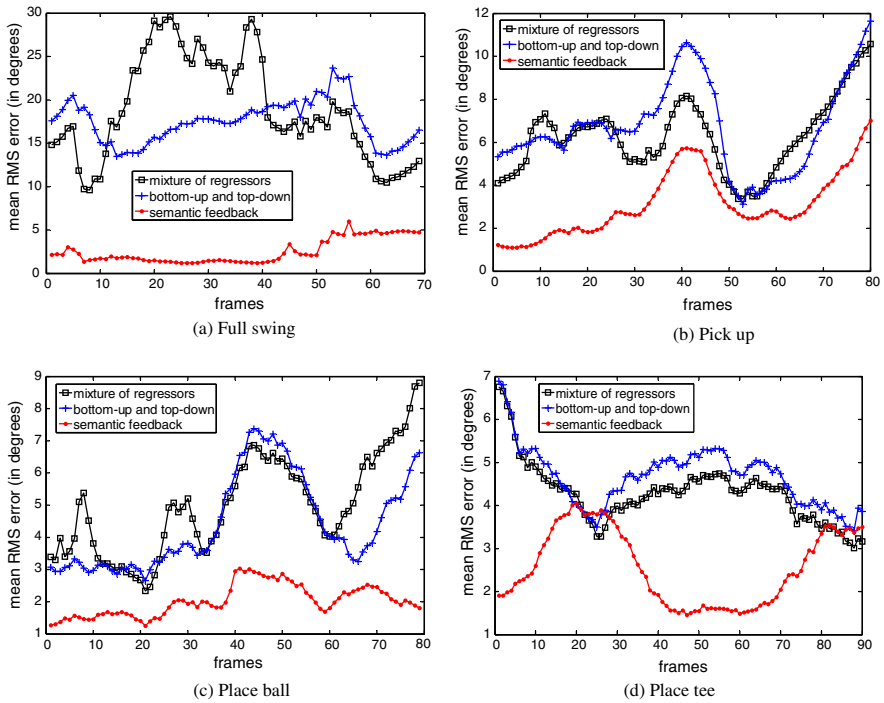

Fig. 8. Comparisons on mean RMS error over testing frames. (a) Full swing. (b) place tee. (d) pick up.(c) place ball. The horizontal axis indicates the frame numbers and the vertical axis indicates the mean RMS errors of all the joint angles in degrees.

Table 1
Comparative results showing RMS errors (average error/maximum joint average error) in degrees.

Motion	Mixture regressors	Bottom-up and top-down	Semantic feedback
Full swing	18.9/29.6	17.5/23.6	**5.1/6.8**
Pick up	7.2/14.7	7.5/14.8	**3.8/9.7**
Place ball	5.6/7.4	5.0/8.8	**2.0/3.0**
Place tee	4.2/6.8	4.5/7.6	**2.6/4.0**

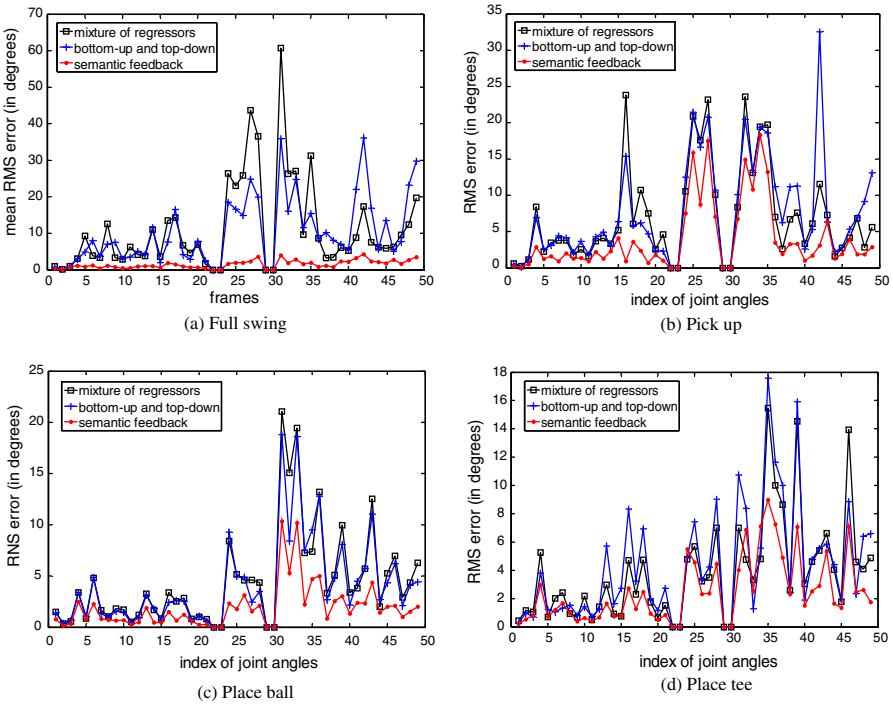

Fig. 9. Quantitative comparisons on the estimation error of joint angles. (a) Full swing, (b) pick up, (c) place the ball and (d) place the tee. The horizontal axis indicates joint angles and the vertical axis indicates the mean RMS errors over the testing sequence in degrees.

method utilizes action recognition feedback for incorporating motion knowledge to alleviate the ambiguity problem. We demonstrate the quantitative comparisons on joint angle average errors in Fig. 9. For the most joint angles, our method has better performance on estimation accuracy with lower RMS error than other methods without recognition feedback.

7. Conclusions and future work

We have presented an action recognition feedback-based framework for articulated human pose reconstruction from monocular images. As action recognition feedback, the high-level motion knowledge is represented by both local spatial motion correlation and global temporal manifold motion template, which are integrated in tandem to handle the ambiguity problem. The local motion correlation represents the constraints between different body parts and the global motion template preserves the temporal relationship between complete body poses. Experiments on the CMU Mocap database illustrate that top-down action recognition feedback can compensate the inadequate appearance cue by exploiting motion knowledge. The future developments of our work will focus on exploiting more motion knowledge such as the self-occlusion constraints between different body parts, style and view model of different human appearances. We are also interested in extending the idea of feeding back high-level knowledge to other visual problems like multiple object tracking and group action analysis.

Acknowledgements

This work was partially supported by the Natural Science Foundation of China (60675021), the Chinese High-Tech Program (2009AA01Z323), and Beijing key discipline program.

References

Agarwal, A., Triggs, B., 2004a. Learning to track 3D human motion from silhouettes. In: Proc. of the Internat. Conf. on Machine Learning. Banff, Canada, pp. 9–16.

Agarwal, A., Triggs, B., 2004b. Tracking articulated motion using a mixture of autoregressive models. In: Proc. of Eur. Conf. on Computer Vision. Prague, Czech Republic, pp. 54–65.

Agarwal, A., Triggs, B., 2005. Monocular human motion capture with a mixture of regressors. In: IEEE Workshop on Vision for Human Computer Interaction.

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

Elgammal, A., Lee, C.S., 2004. Inferring 3D body pose from silhouettes using activity manifold learning. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, vol. 2, Washington, DC, USA, pp. 681–688.

He, X.F., Cai, D., Yan, S.C., Zhang, H.J., 2005. Neighborhood preserving embedding. In: Proc. of IEEE Internat. Conf. on Computer Vision, vol. 2, Beijing, China, pp. 1208–1213.

Hou, S.B., Galata, A., Caillette, F., Thacker, N., Bromiley, P., 2007. Real-time body tracking using a Gaussian process latent variable model. In: Proc. of IEEE Internat. Conf. on Computer Vision, Rio de Janeiro, Brazil.

Huttenlocher, D., Klanderman, G., Rucklidge, W., 1993. Comparing images using the Hausdorff distance. IEEE Trans. Pattern Anal. Machine Intell. 15 (9), 850–863.

Moon, K., Pavlovic, V., 2006. Impact of dynamics on subspace embedding and tracking of sequences. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, vol. 1, New York, USA, pp. 198–205.

Rosales, R., Sclaroff, S., 2006. Combining generative and discriminative models in a framework for articulated pose estimation. Internat. J. Comput. Vision 67 (3), 251–276.

Sminchisescu, C., Kanaujia, A., Li, Z.G., Metaxas, D., 2005. Discriminative density propagation for 3D human motion estimation. In: Proc. of IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, vol. 1, San Diego, CA, USA, pp. 390–397.

Sminchisescu, C., Kanaujia, A., Metaxas, D., 2006. Learning joint top-down and bottom-up processes for 3D visual inference. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, New York, USA, pp. 1743–1752.

Stauffer, C., Grimson, W.E.L., 1999. Adaptive background mixture models for real-time tracking. In: Proc. of IEEE Computer Vision and Pattern Recognition, vol. 2, Santa Barbara, CA, pp. 246–252.

Thayananthan, A., Navaratnam, R., Stenger, B., Torr, P.H.S., Cipolla, R., 2006. Multivariate relevance vector machines for tracking. In: Eur. Conf. on Computer Vision, Prague, Czech Republic, pp. 124–138.

Tipping, M.E., 2001. Sparse Bayesian learning and the relevance vector machine. J. Machine Learn. Res. 1, 211–244.

Wang, J.M., Fleet, D.J., Hertzmann, A., 2008. Gaussian process dynamical models for human motion. IEEE Trans. Pattern Anal. Machine Intell. 20 (2), 282–298.

Xu, X.Y., Li, B.X., 2007. Learning motion correlation for tracking articulated human body with a Rao-Blackwellised particle filter. In: Proc. of IEEE Internat. Conf. on Computer Vision, Rio de Janeiro, Brazil.