Real-time Pose Regression with Fast Volume Descriptor Computation

Michiro Hirai, Norimichi Ukita, Masatsugu Kidode
Nara Institute of Science and Technology
ukita@is.naist.jp

Abstract

We present a real-time method for estimating the pose of a human body using its 3D volume obtained from synchronized videos. The method achieves pose estimation by pose regression from its 3D volume. While the 3D volume allows us to estimate the pose robustly against self occlusions, 3D volume analysis requires a large amount of computational cost. We propose fast and stable volume tracking with efficient volume representation in a low dimensional dynamical model. Experimental results demonstrated that pose estimation of a body with a significantly deformable clothing could run at around 60 fps.

1. Introduction

Analyzing a human body motion from a video(s) have many applications, such as human computer interaction, CG animation, and humanoid robot control.

Marker-less vision-based human motion capture has been proposed in many studies[8]. Previously, a human pose can be estimated mainly from a single viewpoint. However, multi-view analysis has several advantages over single image analysis (e.g. robustness against self occlusions). Recently, shape-from-silhouette (SFS) can compute the 3D volume of a moving person stably in real-time[4]. If 3D volume analysis can be achieved as well as 3D reconstruction, not only precise pose estimation but also online applications using the 3D shape and its motion can be also realized.

In [6], video-rate (30fps) human pose estimation from the 3D volume has been proposed, in which the pose is estimated so that the overlap between the reconstructed volume and a 3D human model consisting of rigid parts is maximized (see [7], for example). Such rigid-parts matching methods cannot be applied to a person who wears loose-fitting clothing (e.g. Japanese kimonos and skirts). While other matching based methods can acquire the pose of such a person by analyzing deformation of clothing as well as the human body (see [2], for example), none of them can work in real time.

In contrast to rigid-part matching based methods, pose-regression based approaches are superior in terms of computational cost and applicability to loose-fitting clothing analysis. The pose regression from image features (e.g. silhouettes and volumes) estimate the human pose based on training data of synchronized pose data and image features, which are used for obtaining regression functions. While fast pose regression has been already achieved (e.g. over 200fps[3]) in 2D image analysis, real-time pose regression from 3D volumes has not been studied much; a successful existing method[11] achieves 3fps in pose regression from 3D volumes.

In this paper, we propose real-time human pose regression from the 3D volume.

2. Related Work

In recent pose estimation works[1], prior knowledge of human motion (motion prior) has been used for precise and robust estimation. Pose tracking using motion prior can resolve ambiguity between image features and the human pose, such as self occlusions.

Particle filter is also popular and powerful for pose tracking. Although it can be done in a relatively low-dimensional joint space (30-D) [5], it is difficult in a more high-dimensional volume space (100-D or more) to distribute particles well. In a low-dimensional latent space obtained by dimensionality reduction (e.g. PCA), particles can be distribute well. Dimensionality reduction can also achieves a low computational cost.

To model complicated human motion in the low-dimensional space, non-linear dimensionality reduction methods have proposed. Among them, Gaussian Process Dynamical Models (GPDM[13]), which provides motion prior (i.e. motion prediction functions) as well as nonlinear probabilistic embedding, is widely used for motion tracking (see [12], for example). Therefore, we employ GPDM for modeling volumetric dynamics in order to track the human volumes for pose regression.

3. Volume Tracking with Dynamical Models

GPDM provides us two mappings. 1) smooth mapping from a point at time $t-1$ to t in the latent space. 2) mapping from volume latent space to volume space. These mappings are modeled as follows:

$$\boldsymbol{x}_t = \sum_i \boldsymbol{a}_i \phi_i(\boldsymbol{x}_{t-1}) + n_{x,t}, \quad (1)$$

$$\boldsymbol{v}_t = \sum_j \boldsymbol{b}_j \psi_j(\boldsymbol{x}_t) + n_{v,t}, \quad (2)$$

where $\boldsymbol{v}_t \in \Re^d$ is a zero-mean volume data at time t, $\boldsymbol{x}_t$ is the respective latent variable ($d \ll D$); in this paper, $D = 180$ in the volume descriptor (will be introduced in Sec. 4) and $d = 6$ in its latent space. ϕ_i and ψ_k are basis functions with weights $\boldsymbol{\mathcal{A}} = [\boldsymbol{a}_1, \cdots]$, $\boldsymbol{\mathcal{B}} = [\boldsymbol{b}_1, \cdots]$, and $n_{x,t}, n_{v,t}$ denote noise. Under the assumption that the noise is zero-mean Gaussian, the likelihood of $\boldsymbol{V} = [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_N]^T$ (denoted by $p(\boldsymbol{V}|\boldsymbol{X}, \alpha)$, where $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N]$ and α is hyper parameters) can be obtained by marginalizing over $\mathcal{B}$.

Similarly, the likelihood for $\boldsymbol{X}$, $p(\boldsymbol{X}|\beta)$ where β is hyper parameters, can be obtained from Eq. (1).

Volumetric Dynamical Model Learning

Learning the GPDM entails estimation the latent variables $\boldsymbol{X}$ and the hyperparameters $[\alpha, \beta]$. Following [13] we adopt simple prior distributions over hyperparameters, GPDM posterior becomes

$$p(\boldsymbol{X}, \alpha, \beta|\boldsymbol{V}) \propto p(\boldsymbol{V}|\boldsymbol{X}, \beta)p(\boldsymbol{X}|\alpha)p(\alpha)p(\beta), \quad (3)$$

whose log posterior is maximized by Scaled Conjugate Gradient. The acquired $\boldsymbol{X}$ is shown in Fig. 1. In this figure, a color variation shows the variances in each location in X; low (red) to high (blue) variances.

Mapping Functions

With the optimized $\boldsymbol{X}$, the mapping function from X to V and the temporal mapping function can be obtained. The variance (i.e. inverse likelihood) of the mapping function from X to V is also given. See the original paper[13] for the details.

A mapping function $V \rightarrow X$ is not explicitly provided by GPDM. In our method, this mapping is obtained by GP regression from $\boldsymbol{V}$ to $\boldsymbol{X}$.

Latent Volume Tracking

With the mapping functions mentioned above, volume tracking is achieved. Volume tracking is implemented in particle filtering in the latent space X.

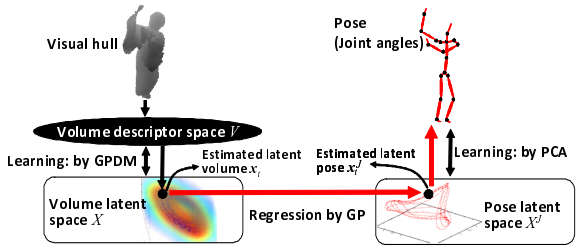

Figure 1. Pose regression from volume.

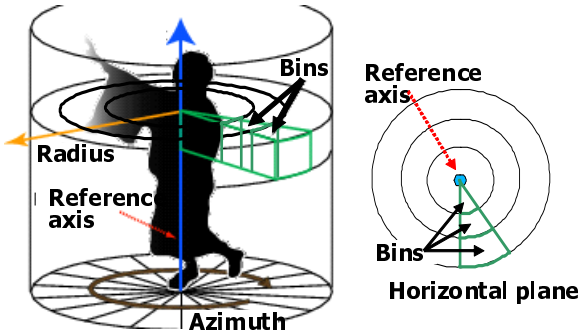

Figure 2. Structure of a volume descriptor.

An input volume data $\boldsymbol{v}_t$ at each capturing timing t is rotated θ, which is the azimuth difference between $\boldsymbol{v}_{t-1}$ and each particle at $t-1$, for azimuth alignment and mapped to X. The likelihood of each particle $\boldsymbol{x}_i^p$ with respect to the mapped volume $\boldsymbol{x}_t^\theta$ is then computed while θ is changed as follows:

$$\max_\theta \left(\exp\left(-\frac{\|\boldsymbol{x}_t^\theta - \boldsymbol{x}_i^p\|^2}{\nu} \right) \exp\left(-\frac{\sigma_V^2(\boldsymbol{x}_t^\theta)}{w} \right) \right) \quad (4)$$

where, ν and w denote weight parameters and $\sigma_V^2(\boldsymbol{x})$ is the variance of the mapping function from X to V. Finally, the likelihood-weighted mean of the particles ($\bar{\boldsymbol{x}}_t$ in Fig. 1) is regarded as the tracking result at t.

4. Fast Volume Descriptor Computation

In our method, the volume data $\boldsymbol{v}$ processed by GPDM is a *volume descriptor*. Our volume descriptor is a modified version of a voxel data descriptor[9]. The volume descriptor is used for improving robustness to reconstruction noise and dimensionality reduction, which allows GPDM to optimise temporal volumes. As proposed in [11], the volume descriptor is reconstructed from a visual hull and rotated in azimuth alignment with training volume data.

A voxel set is reconstructed by SFS from synchronized images. The perpendicular axis drawn through the centroid of the voxels is regarded as a reference axis

of the volume descriptor. The voxels are divided into fan-shaped bins as illustrated in Fig. 2: height division n_z, azimuth division n_a, and radius division n_r, based on the reference axis. The 3D positions of each voxel $\boldsymbol{v}_i = [x_i, y_i, z_i]$ in the world coordinate system is convert to $\boldsymbol{p}_i = [z_i, a_i, r_i] = [\sqrt{x_i^2 + y_i^2}, \mathrm{atan}(y/x), z_i]$ in the cylindrical polar coordinates. The volume descriptor is computed by counting the number of the surface voxels in every bin; the volume descriptor is a $n_z \times n_a \times n_r$-dimensional vector.

As described in Sec. 3, an input volume at each moment must be rotated for azimuth alignment with every particle. This needs a large amount of computational cost in the previous work[11].

With 3D coordinates of the surface voxels within the bins in the cylindrical polar coordinates, only three arithmetic operations for each surface voxel are needed for computing the rotated volume descriptor. For example, rotating the input volume ψ degrees is expressed by $\tilde{\boldsymbol{p}}_{\boldsymbol{i} \in \boldsymbol{vol}} = [z_i, a_i + \psi, r_i]$. Therefore, rotation is calculated by only one adding operation. Then the azimuth bin-index of each surface voxel is computed by $\lfloor (a_i + \psi)/n_a \rfloor$. Experimental results show that volume conversion and rotation worked within 0.015 sec using 1^3 cm^3 voxels within a 165 cm height voxel set.

5. Pose Regression from Volume

In offline learning, the pose latent space X^J is also generated from sample poses (i.e. a set of joint angles $\boldsymbol{J} = [\theta_1, \cdots, \theta_{N^J}]$, where N^J denotes the number of the joints), each of which is synchronized with its sample volume $\boldsymbol{v}$, by PCA. Then pose regression function from X to X^J is learned by GP.

In online pose tracking, volume tracking is first achieved as introduced in Sec. 3. From the volume tracking result $\hat{\boldsymbol{x}}_t$, the respective pose latent variable $\boldsymbol{x}_t^J$ is obtained (red right-pointing arrow in Fig. 1). Finally, the pose at t is estimated from back-projection of $\boldsymbol{x}_t^J$ (red upward arrow in Fig. 1).

6. Experimental Results

In each sequence, a subject who wore loose-fitting clothing was captured by 8 roof-mounted synchronized cameras at 30 fps (1024×768 pixels). Two kinds of dance sequences (dance1 and dance2) were captured.

For each kind of dance, 350-frame temporal volumes, reconstructed by [10], were used for learning the model. In this sample volume sequence, only one subject was observed. 300-frame test sequences were prepared by observing three subjects.

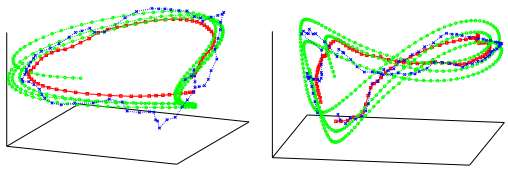

Figure 4. Tracking results in the volume latent spaces (green: learned latent model, blue: input visual hulls, red: tracking results). Left: dance1, Right: dance2.

To obtain pose data for training data and evaluation, a motion capture system (IGS-190) was used. With this motion capture system, 54 dimensional pose data (3-DOF × 18 joints) was obtained.

We empirically determined the dimensions of the volume descriptor and the volume latent space X: $[n_z, n_\theta, n_r] = [5, 16, 2]$ (180-D volume descriptor) and the 6-D volume latent space. The dimension of pose latent space was determined so that its cumulative ratio of the covariance matrix of training pose data was over 0.95; the dimension of X^J was 4.

Figure 3 shows the results of pose estimation of dance1. It can be seen that the estimated poses were visually reasonable.

Figure 4 shows the volume tracking results in low dimensional latent spaces of the volume descriptors. While the trajectories of input visual hulls were zig-zag and away from the learned models due to reconstruction errors, tracking with the prediction model and particle filtering was smooth and close to the learned models.

Our pose tracking ran around 60 fps. This computational cost is 20 times faster than that of the previous work of pose regression from volume[11].

For accuracy evaluation, the mean of RMS errors (i.e. the mean of all subjects, all joints, and all frames) of the absolute difference between the estimated joint angles and the respective ground truth was evaluated. While the mean errors were 7.3 and 8.8 degrees (in dance1 and dance2 sequences, respectively) in the proposed method, those by the previous method[11] were 5.0 and 5.0 degrees. The accuracy was declined in contrast to the previous method because it employs an additional constraint for volume likelihood computation, which is computationally heavy but useful for accurate likelihood computation.

7. Concluding Remarks

With efficient volume descriptor computation, the proposed method runs at over the video rate (i.e. 60fps),

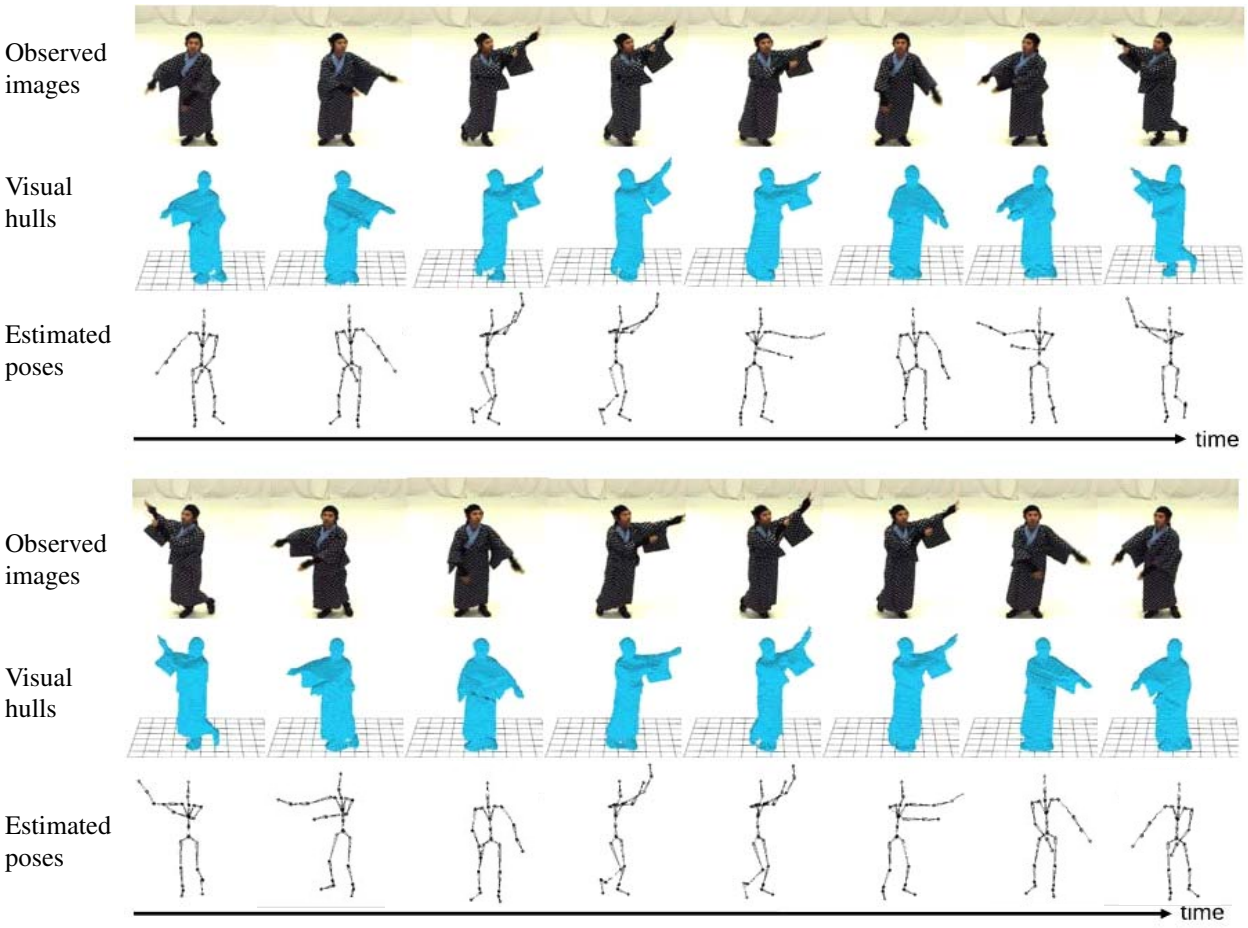

Figure 3. Pose estimation results of a dance1 sequence.

which is enough for online applications. In contrast to the original work[11], the computational cost in rotating volume particles is reduced by simplified arithmetic operations with the cylindrical polar representation of the volume.

Future work includes improving 1) robustness and accuracy of pose estimation and 2) scalability of motion prior learning.

References

[1] A. Agarwal and B. Triggs. Tracking articulated motion using a mixture of autoregressive models. In *ECCV*, 2004.

[2] A. O. Balan and M. J. Black. The naked truth: Estimating body shape under clothing. *ECCV*, 2008.

[3] A. Bissacco, M.-H. Yang, and S. Soatto. Fast Human Pose Estimation using Appearance and Motion via Multi-Dumensional Boosting Regression. *CVPR*, 2007.

[4] G. Cheung, T. Kanade, J. Bouguet, and M. Holler. A Real Time System For Robust 3d Voxel Reconstruction of Human Motions. *CVPR*, 2000.

[5] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. In *CVPR*, 2000.

[6] S. Hou, A. Galata, F. Caillette, N. Thacker, and P. Bromiley. Real-time Body Tracking Using a Gaussian Process Latent Variable Model. In *ICCV*, 2007.

[7] I. Mikić, M. Trivedi, E. Hunter, and P. Cosman. Human Body Model Acquisition and Tracking Using Voxel Data. *IJCV*, 53(3):199–223, 2003.

[8] R. Poppe. Vision-based Human Motion Analysis: An Overview. *CVIU*, 108(1-2):4–18, 2007.

[9] Y. Sun, M. Bray, A. Thayananthan, B. Yuan, and P. H. S. Torr. Regression-based human motion capture from voxel data. In *BMVC*, 2006.

[10] T. Tung, S. Nobuhara, and T. Matsuyama. Simultaneous super-resolution and 3d video using graph-cuts. In *CVPR*, 2008.

[11] N. Ukita, M. Hirai, and M. Kidode. Complex volume and pose tracking with probabilistic dynamical models and visual hull constraints. In *ICCV*, 2009.

[12] R. Urtasun, D. Fleet, A. Hertzmann, and P. Fua. 3d people tracking with gaussian process dynamical models. In *CVPR*, 2006.

[13] J. Wang, D. Fleet, and A. Hertzmann. Gaussian process dynamical models for human motion. *PAMI*, 2007.