Complex Volume and Pose Tracking with Probabilistic Dynamical Models and Visual Hull Constraints

Norimichi Ukita†‡ Michiro Hirai† Masatsugu Kidode†
†Nara Institute of Science and Technology
‡The Robotics Institute, Carnegie Mellon University
ukita@ieee.org

Abstract

We propose a method for estimating the pose of a human body using its approximate 3D volume (visual hull) obtained in real time from synchronized videos. Our method can cope with loose-fitting clothing, which hides the human body and produces non-rigid motions and critical reconstruction errors, as well as tight-fitting clothing. To follow the shape variations robustly against erratic motions and the ambiguity between a reconstructed body shape and its pose, the probabilistic dynamical model of human volumes is learned from training temporal volumes refined by error correction. The dynamical model of a body pose (joint angles) is also learned with its corresponding volume. By comparing the volume model with an input visual hull and regressing its pose from the pose model, pose estimation can be realized. In our method, this is improved by double volume comparison: 1) comparison in a low-dimensional latent space with probabilistic volume models and 2) comparison in an observation volume space using geometric constrains between a real volume and a visual hull. Comparative experiments demonstrate the effectiveness of our method faster than existing methods.

1. Introduction

With human pose information, a number of real-world applications can be realized. Such techniques have been proposed in many studies[1]. To estimate complex dynamic poses, pose tracking using motion prior and multiview images is more effective than pose detection from a unidirectional view. Although a number of works using motion have been proposed (e.g., [2, 3]), pose estimation using multiview images has not been studied extensively. This is because multiview analysis (i.e., 3D reconstruction) requires a large amount of computational time and might not be robust against 3D errors. For example, while latest stereo algorithms such as [4] can reconstruct a detailed shape, its computational cost is expensive. However, recent Shape-From-Silhouette (SFS) can compute the volume (i.e., visual hull) of a person moving stably in real time [5, 6].

In most methods using 3D volumes, the pose is estimated so that the overlap between the reconstructed volume and a 3D human model that consists of *simple rigid* parts (e.g., cylinders) is maximized (e.g., [7, 8]). These methods can work well under the assumption that each body part is approximately modeled as a rigid part. This assumption, however, cannot represent *a large variation* of *non-rigid loose-fitting* clothing.

In [9], the shapes of a skirt and legs are reconstructed simultaneously. However, it is very slow (5min/frame) and can cope only with simple shape variations (i.e., simple pose of only almost observable legs). With [10], a body shape under clothing can be reconstructed. Although the pose can be estimated from the body shape, 1) a large part of limb should be observable, 2) deformation of loose-fitting clothing makes reconstruction difficult, and 3) reconstruction is very slow (40min/frame) in [10]. On the other hand, Ukita et al.[11] proposed real-time body-part identification in the volume of a person who wears loose-fitting clothing. However, it cannot estimate the pose itself. Furthermore, it cannot cope with complex shape deformation because its volume model is simple. As far as we know, no existing algorithm can achieve pose estimation of a human body *hidden by loose-fitting clothing* in *complex* motion. In this paper, to cope with this kind of difficult motion, we propose novel pose tracking using temporal volumes and probabilistic human motion dynamics.

2. Related Work

The pose of a human body is modeled by a set of joint angles. For pose estimation from an image(s), some kind of shape feature (e.g., boundary line, silhouette, volume) that expresses a human body shape is extracted. Then the pose is estimated based on the geometric correspondence between the body shape and the pose. Prior knowledge of a

2009 IEEE 12th International Conference on Computer Vision (ICCV)
978-1-4244-4419-9/09/$25.00 ©2009 IEEE

human body (e.g., structure and motion) can improve accuracy and robustness of pose estimation. The more detailed and precise the prior becomes, the more accurate and robust its estimation result gets. The detailed and precise prior of the human body can be obtained by a Motion Capture (MoCap) system. While several kinds of body information are obtained with MoCap, most works focus on the motion of the joint angles: motion models with Gaussian mixture models[12], HMM[13], and autoregressive models[3]. The motion prior is useful for resolving the short-lasting ambiguities between a body shape and its pose.

However, the high dimensionality of the joint angles (30-60 dimensions) and their erratic motions make it difficult to represent various motions efficiently and correctly. Therefore, their motion prior is expressed probabilistically (e.g., by using HMM/Gaussian) in a lower dimensional space (e.g., by using PCA). Recently, a series of Gaussian Process Latent Variable Models (GPLVM[14]), which provides nonlinear probabilistic embedding, is widely used for motion learning: dynamics representation[15], bidirectional smooth mapping[16], and shared latent structure[17]. This kind of embedding is also useful for modeling high-dimensional shape features such as silhouettes and volumes as proposed in [17] and [18], respectively.

Particle filtering is also popular for pose tracking. Although it can be done in a high-dimensional joint space by using particle reposition[19] and/or coarse-to-fine processing[20], it is difficult in a more high-dimensional space (100-D or more); in our framework, the particle dimension is 160-D, in which it is impossible to distribute particles well. Therefore, the above embedding is useful for getting feasible dimensional variables.

In general, joint angles are measured by an optical MoCap system, in which markers are attached on a human body. The optical MoCap, however, cannot observe the body if it is hidden by loose-fitting clothing such as a skirt or dress. Even for such a target, a MoCap with accelerometers and/or geomagnetism magnetism sensors can measure the joint angles (e.g., off-the-shelf product[21]).

3D shape reconstructed from multiview images can also improve pose estimation under occlusion as described in Sec. 1. With 3D comparison, the pose that is ambiguous in the silhouette is also estimated correctly.

None of the above methods, however, can estimate the pose of a person who wears significantly deforming loose-fitting clothing because they assume rigid articulated motion. To cope with this problem, a regression based approach is useful. In its learning process, each pose data is obtained and recorded with its corresponding shape data (e.g., silhouette/volume). In its estimation process, then, the pose of an input shape data is inferred from the input by regression. In [22, 17], regression from the 2D silhouette is achieved by RVM and shared-structure GPLVM, respectively. In these methods, efficient and robust shape features[23] extracted from each silhouette are used for regression. As well as regression from the silhouette, regression from the volume has been also proposed[24, 18]. In [24, 18], 3D extension[25] of the above shape feature is employed as in [26]. This kind of 3D shape representation has been studied in 3D retrieval; see review[27], for example. These features are efficient and still robust against noise. However, none of the previous works can be applied to estimating the pose of a person wearing loose-fitting clothing. This is because ambiguity between the human and clothing shape and its pose increases significantly due to large deformation of clothing; even if the observable clothing shapes are the same, the unobservable body poses have a large variation depending on previous motions.

Furthermore, while SFS is fast and stable, it is not easy to estimate the pose from the visual hull reconstructed by SFS even with regression. This is because the visual hull may include large errors mainly in concave regions of the human body and clothing. We call these errors **phantom volumes**. Note that they change depending not only on the shape of a target but also on the geometric configuration (location and orientation) among the target and cameras. This results in difficulty in volume matching for regression. This is one of the major problems not only in pose regression but also in all kinds of pose estimation from the visual hull. The phantom volumes can be refined based on post-processes such as multi-view photo consistencies (e.g., dense stereo[4] space carving[28]) and additional restrictions such as silhouette consistencies and spatio-temporal smoothness (e.g., the deformable mesh model[29]). In [30], a template human model is used for more reliable mesh deformation. These methods, however, require a large amount of computational cost for 3D reconstruction (e.g., one minute or more).

3. Basic Scheme for Probabilistic Volume and Pose Tracking

In our *offline learning scheme* (enclosed by a thick red solid line in Fig. 1), training samples of synchronized body volumes and their poses are obtained. The sample volume is reconstructed by a slow but sophisticated algorithm[4, 28, 29] that produces few errors. That is, the sample volumes are not visual hulls but correctly reconstructed volumes. We call them **refined volumes**. The volume model is, therefore, invariant to the geometric configuration between the target person and cameras unlike the visual hull. Learning not the visual hulls but the refined volumes has one more advantage; in the visual hulls, the difference between the shapes might be buried in the phantom volumes. For efficient and robust features extracted from the volume, we improves a 3D model proposed in [24, 18]. We call this model a **volume descriptor** (described in Sec. 4), which is designed to fit our volume track-

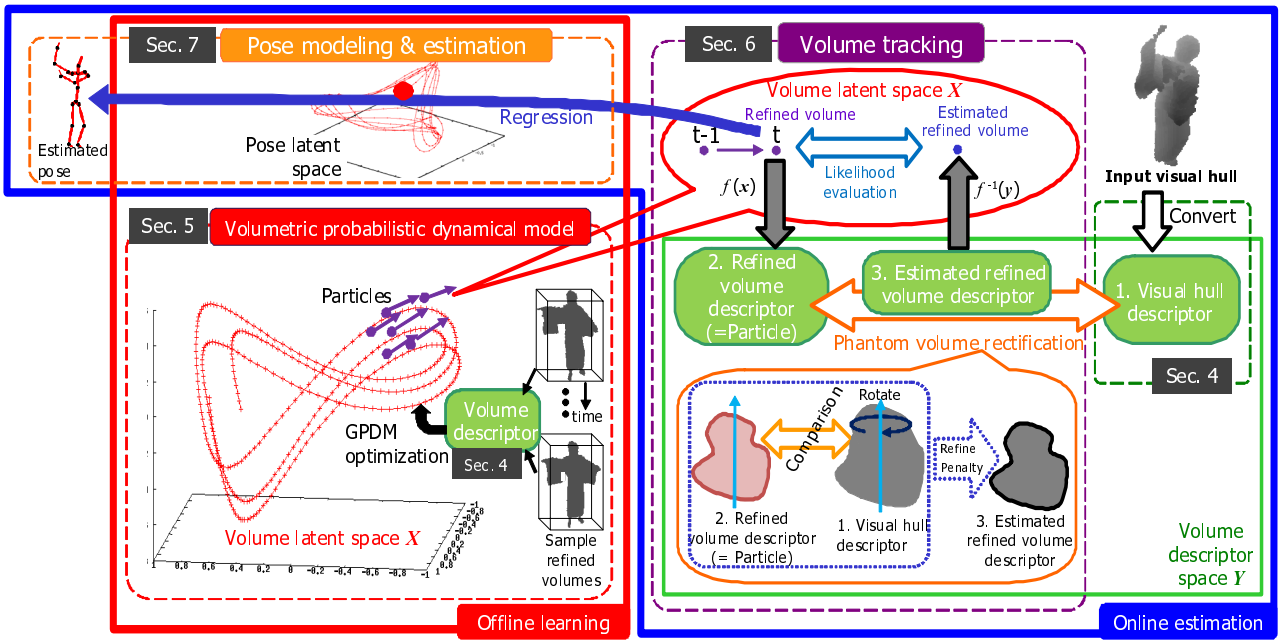


Figure 1. Algorithm overview. Each of dashed lines encloses the contents described in each section.

ing scheme. From the volume descriptors of all the sample refined volumes, their probabilistic dynamical model in a lower-dimensional latent space (as shown in "Volume latent space X" in Fig. 1) is acquired by Gaussian Process Dynamical Models[15] (GPDM) as shown in "GPDM optimization" in Fig. 1 (described in Sec. 5). With this model, bidirectional mappings between the observation space, Y, and the latent space, X, are obtained (denoted by $f(\boldsymbol{x})$ for $X \rightarrow Y$ and $f^{-1}(\boldsymbol{y})$ for $Y \rightarrow X$ in Fig. 1, respectively). The pose space is also modeled as a lower-dimensional latent space. Finally, the mapping from the volume latent space to the pose latent space is learned using RVM[31].

Our *online pose tracking scheme* (enclosed by a thick blue solid line in Fig. 1) consists of volume tracking (described in Sec. 6) and pose regression from the volume (described in Sec. 7). In volume tracking, particle filtering in X is achieved. Each particle corresponds to a refined volume descriptor in X. At each frame, an input visual hull ("Input visual hull" shown in the top right in Fig. 1) is reconstructed by SFS and converted to its volume descriptor ("1. Visual hull descriptor" in Fig. 1). In a simple implementation of particle filtering, the visual hull descriptor is mapped by $f^{-1}(\boldsymbol{y})$ into X and then the likelihood between each particle and the mapped visual hull descriptor is evaluated. This comparison is, however, irrelevant because the particle is computed from the refined volumes while the input is computed from the visual hull. This problem did not emerge in previous works because loose-fitting clothing, which makes phantom volumes, was not used. In the 3D real-world space, the refined volume must be geometrically encapsulated in the visual hull. We call this constraint the **visual hull constraint**, which is crucial in our pose tracking. To evaluate the visual hull constraint, the input visual hull is compared with the refined volumes in the volume descriptor space ("Phantom volume comparison" in Fig. 1), in which this constraint can be evaluated efficiently. For that, each particle is mapped to Y (i.e., "2. Refined volume descriptor" in Fig. 1) by $f(\boldsymbol{x})$. The input visual hull modified by this comparison (i.e., "3. Estimated refined volume descriptor" in Fig. 1) is then mapped into X and then compared with all the particles for computing their likelihoods. Finally, the pose is estimated from the particles and their likelihoods (i.e., the likelihood-weighted mean of the particles) by regression learned using RVM.

4. Efficient Volume Descriptor

Our volume descriptor is a modified version of voxel data description[24, 18]. In [24, 18], voxel data is divided into several bins. In each bin, the number of the surface voxels in the reconstructed volume is counted. The shape of the bin model of our descriptor is a cylinder, whose center is a reference axis (Fig.2). The median of all human voxels is regarded as the reference axis. The size of the cylinder in our experiments was as follows: height = $1.23H$ and radius = $0.6H$, where H is the body height of a subject, for normalizing the subjects' heights. While the bin model in [24] has height, radius, and azimuth divisions, our bin model has only height and azimuth divisions. Instead

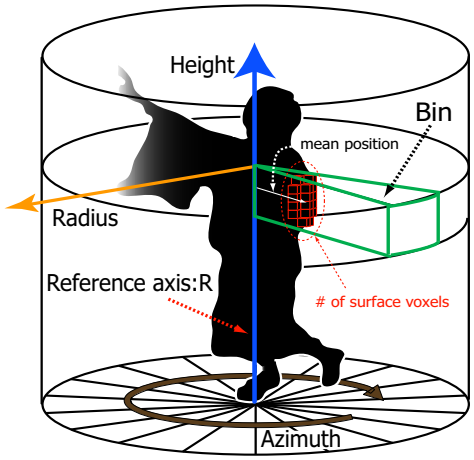


Figure 2. Bin structure for our volume descriptor.

of radius division, our descriptor has not only the number of surface voxels but also their mean position (i.e., a mean distance from the reference axis) in each bin. The dimension of our descriptor with N_h height divisions and N_a azimuth divisions is decreased to $\frac{2}{N_r}$, where N_r is radius divisions, of that in [24]. This efficient dimensionality reduction improves GPDM optimization. This results in high neighborhood-preserving property and invertibility of the mapping functions (i.e., $y = f(x)$ and $x = f^{-1}(y)$). The former property is important for correct motion prior in the latent space. The high invertibility is indispensable for our visual hull constraint. The mean position along the radius axis instead of the radius division is suitable also for the visual hull constraint (see Sec. 6).

5. Volumetric Probabilistic Dynamical Model

GPDM[15] provides us two mappings, 1) from a point at $t-1$ to a point at t in the latent space and 2) from a latent space to an observation space. In particular, the former mapping is useful for volume tracking. These two mappings are modeled as follows:

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

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

where $\boldsymbol{v}_t$ is a D-dimensional volume descriptor at time t, $\boldsymbol{x}_t$ is its d-dimensional latent variable ($d \ll D$), ϕ_i and ψ_j are basis functions with weights $\boldsymbol{A} = [\boldsymbol{a}_1, \dots]$ and $\boldsymbol{B} = [\boldsymbol{b}_1, \dots]$, and $n_{x,t}$ and $n_{v,t}$ are noise. Under the assumption that the noise is zero-mean Gaussian, the following likelihood for $\boldsymbol{V} = [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_N]$ can be obtained by marginalization of the basis functions in Formula (2):

$$p(\boldsymbol{V}|\boldsymbol{X}, \theta) = \frac{1}{\sqrt{(2\pi)^{ND} \|K_V\|^D}} \exp\left(-\frac{1}{2}\mathrm{tr}(\boldsymbol{K}_V^{-1}\boldsymbol{V}\boldsymbol{V}^T)\right), \quad (3)$$

where N is the number of samples, $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N]$, and $\boldsymbol{K}_V$, in which $\boldsymbol{K}_{V_{i,j}} = k_V(\boldsymbol{x}_i, \boldsymbol{x}_j)$, is a kernel matrix with

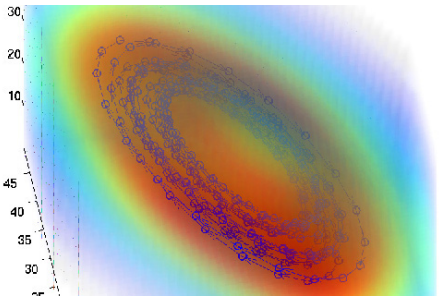


Figure 3. Probabilistic dynamical model. Circles: samples, Arrows: motion, Color: variance (blue(low) → red(high)).

hyperparameters θ. In our experiments, the nonlinear radial basis function was used for the kernel function $k_V(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Similarly, the likelihood for $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N]$ can be obtained from Formula (1):

$$p(\boldsymbol{X}|\theta_X) = \frac{p(\boldsymbol{x}_1)}{\sqrt{(2\pi)^{(N-1)D} \|\boldsymbol{K}_X\|^D}} \exp\left(-\frac{1}{2}\mathrm{tr}(\boldsymbol{K}_X^{-1}\boldsymbol{X}_o\boldsymbol{X}_o^T)\right), \quad (4)$$

where $\boldsymbol{X}_o = [\boldsymbol{x}_2, \cdots, \boldsymbol{x}_N]$ and $\boldsymbol{K}_X$ is a kernel matrix constructed from $[\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{N-1}]$ with hyperparameters θ_X.

Constructing a latent space with $\boldsymbol{X}$ requires optimization of the hyperparameters θ and θ_X for maximizing the joint likelihood of Formulas (3) and (4). The latent space constructed from a sample volume sequence, which was used in our experiments, is shown in Fig. 3. For visualization, its dimension is set to three.

With the optimized latent space, the mean and variance of $\boldsymbol{v}_t$, $\mu_V(\boldsymbol{x})$ and $\sigma_V^2(\boldsymbol{x})$, and the mean of $\boldsymbol{x}_t$, $\mu_X(\boldsymbol{x})$, can be computed from $\boldsymbol{x}_t$ as follows:

$$\mu_V(\boldsymbol{x}_t) = \mu_v + V^T K_V^{-1} k(\boldsymbol{x}_t) = \boldsymbol{v}_t, \quad (5)$$

$$\sigma_V^2(\boldsymbol{x}_t) = k(\boldsymbol{x}_t, \boldsymbol{x}_t) - k(\boldsymbol{x}_t)^T K_V^{-1} k(\boldsymbol{x}_t), \quad (6)$$

$$\mu_X(\boldsymbol{x}_{t-1}) = \mu_x + X_o^T K_X^{-1} k(\boldsymbol{x}_{t-1}) = \boldsymbol{x}_t. \quad (7)$$

$\mu_V(\boldsymbol{x})$ and $\mu_X(\boldsymbol{x})$ are employed for a mapping function from the volume space to the latent space (i.e., $f(\boldsymbol{x}) = \mu_V(\boldsymbol{x})$) and a motion prior function from $\boldsymbol{x}_{t-1}$ to $\boldsymbol{x}_t$, respectively.

A mapping function from the volume space to the latent space ($f^{-1}(\boldsymbol{v}) : \boldsymbol{V} \rightarrow \boldsymbol{X}$) is not explicitly provided by GPDM. It is, however, required for our visual hull constraint in volume tracking. Back-constrained GPLVM[16] gives us the mapping $\boldsymbol{V} \rightarrow \boldsymbol{X}$ as well as the mapping $\boldsymbol{X} \rightarrow \boldsymbol{V}$. However, more constraints (i.e., variables) are needed in its optimization, which is difficult to converge successfully. In this work, the mapping $\boldsymbol{V} \rightarrow \boldsymbol{X}$ is obtained by regression, which is learned by multi-variate RVM[31].

6. Probabilistic Volume Tracking with Visual Hull Constraint

For pose regression from a volume, the correct refined volume is required. In our method, the volume is tracked

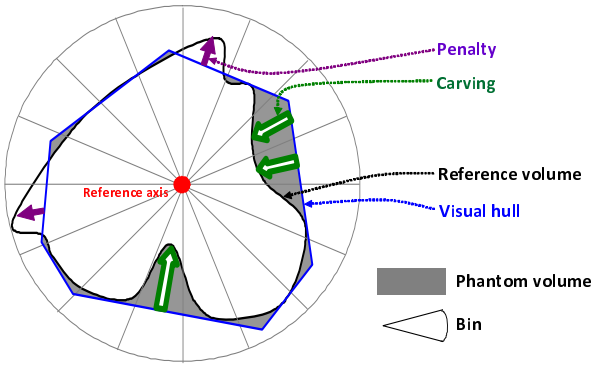


Figure 4. Visual hull constraint in an azimuth-radius plane.

in order to obtain the correct one without being disturbed by phantom volumes of an input visual hull. To achieve this tracking robustly and efficiently, particle filtering in the volume latent space given by GPDM is executed (as shown as "Particles" in "Volume latent space X" in Fig. 1).

In addition, the visual hull constraint is used for comparison between an input visual hull and a refined volume learned in the latent space. In principle, the refined volume is geometrically encapsulated in the visual hull. Therefore, for matching between the visual hull and the refined volume, 1) regions inside the refined volume and outside the visual hull should be penalized (as depicted by arrows "Penalty" in Fig. 4) and 2) regions inside the visual hull and outside the refined volume are the potential regions of phantom volumes (as "Carving" in Fig. 4).

However, our volume tracking is executed in the latent space X, in which the visual hull constraint cannot be evaluated. Each particle (denoted by $\boldsymbol{x}_i^p | i \in \{1, \cdots, N^p\}$, where N^p is the number of the particles) is, therefore, transformed to its volume descriptor by mapping $X \rightarrow Y$, $f(\boldsymbol{x}_i^p)$ and compared with the input visual hull descriptor (denoted by $\boldsymbol{v}$). To evaluate the visual hull constraint in the volume descriptors, their mean positions of surface voxels in b-th bin (denoted by r_b^{vh} for the visual hull descriptor and r_b^{rv} for the refined volume descriptor obtained from a particle, $f(\boldsymbol{x}_i^p)$) are compared with each other as follows:

1) Penalty The penalty $\mathcal{P}_i$ of each particle is determined from the distance between r_b^{vh} and r_b^{rv}:

$$\mathcal{P}_i \quad = \quad \sum_{b=1}^{N_h N_a} p(b) + C^p, \tag{8}$$

where $p(b) = \max(r_b^{rv} - r_b^{vh}, 0)$ and C^p is a constant.

2) Carving The potential regions of phantom volumes in b-th bin, where $r_b^{rv} < r_b^{vh}$, of $\boldsymbol{v}$ are carved so that r_b^{vh} gets closer to r_b^{rv}. It is, however, dangerous to move r_b^{vh} to r_b^{rv} without taking into account the reliability of each volume (denoted by $c(\cdot)$); this may produce over-carving without paying attention to physical limitations (e.g., volume conservation) and the probabilistic volume model (i.e., the volume latent space X). Therefore, a carved r_b^{vh} (denoted by $\hat{r}_b^{vh}$) is obtained taking into account the reliabilities of $f^{-1}(\boldsymbol{v})$ and $\boldsymbol{x}_i^p$ in X, $c(\boldsymbol{x})$, as follows:

$$\hat{r}_b^{vh} \quad = \quad r_b^{vh} - (r_b^{vh} - r_b^{rv})(1 - c(f^{-1}(\boldsymbol{v})))c(\boldsymbol{x}_i^p) \tag{9}$$

$$c(\boldsymbol{x}) \quad = \quad \exp(-\sigma_V^2(\boldsymbol{x})/w), \tag{10}$$

where $\sigma_V(\boldsymbol{x})$ denotes the variance of $\boldsymbol{x}$ (obtained by Formula (6)) in the latent space, w is a weight variable, which might have different values for $c(f^{-1}(\boldsymbol{v}))$ and $\boldsymbol{x}_i^p$. Note that the latent variable of the input visual hull, $f^{-1}(\boldsymbol{v})$, which is not illustrated in Fig. 1 for simplification, must computed for getting its variance in the latent space, namely $c(f^{-1}(\boldsymbol{v}))$.

With the visual hull constraint, our volume tracking at each frame is designed as follows:

1. A visual hull is reconstructed by SFS.

2. The reliability of i-th particle, $c(\boldsymbol{x}_i^p)$, is computed by Formula (10). All particles are then mapped to their volume descriptors, $f(\boldsymbol{x}_i^p)$, by Formula (5).

3. An input visual hull must match with $f(\boldsymbol{x}_i^p)$, whose shape is similar to that of the input but its azimuth orientation differs from that of the input. For this matching, 1) the input visual hull is rotated by θ, 2) its volume descriptor $\boldsymbol{v}^\theta$ is computed, and 3) the orientation of the input, $\hat{\theta}$, is estimated by finding θ, which maximizes the following weighted-sum of the normalized cross-correlation between $\boldsymbol{v}^\theta$ and $f(\boldsymbol{x}_i^p)$: $\sum_i c(\boldsymbol{x}_i^p) \frac{\boldsymbol{v}^\theta f(\boldsymbol{x}_i^p)}{\sqrt{((\boldsymbol{v}^\theta)^2 (f(\boldsymbol{x}_i^p))^2)}}$.

4. The input visual hull rotated by $\hat{\theta}$ is converted to its volume descriptor, $\boldsymbol{v}^{\hat{\theta}}$.

5. With comparison with $f(\boldsymbol{x}_i^p)$, $\boldsymbol{v}^{\hat{\theta}}$ is carved by Formula (9) and its penalty is computed by Formula (8). The carved $\boldsymbol{v}^{\hat{\theta}}$ is then mapped to the latent space.

6. Let $\boldsymbol{x}_i^{\hat{\theta}}$ denote the latent variable of $\boldsymbol{v}^{\hat{\theta}}$ carved by $f(\boldsymbol{x}_i^p)$. The likelihood of $\boldsymbol{x}_i^{\hat{\theta}}$ for $f(\boldsymbol{x}_i^p)$ is expressed as follows: $\exp(-\frac{||\boldsymbol{x}_i^{\hat{\theta}} - \boldsymbol{x}_i^p||^2}{\nu}) c(\boldsymbol{x}_i^p) \mathcal{P}_i^{-1}$, where ν is a weight variable that determines the weight of $||\boldsymbol{x}_i^{\hat{\theta}} - \boldsymbol{x}_i^p||$ in the likelihood computation.

7. The mean of all particles, each of which is weighted by its likelihood, is regarded as the latent variable of the current volume.

8. All particles are shifted temporally with Formula (7).

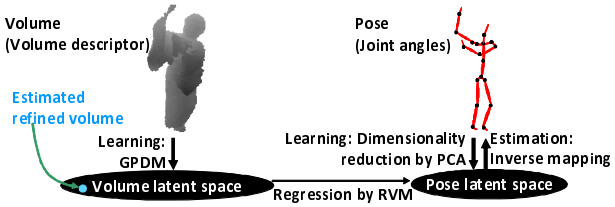


Figure 5. Pose regression from the estimated refined volume.

7. Pose Regression from the Estimated Volume

Figure 5 shows the overview of pose regression from the refined volume estimated by volume tracking. In the learning process, the eigenspace is generated from all samples of joint angles, which are obtained in synchronization with the sample volumes. This eigenspace is regarded as the pose latent space. Then pose regression learning is achieved by multivariate RVM[31]. In the pose estimation processing, the latent variable of a pose is regressed from the estimated refine volume ("Estimated refined volume" in Fig. 5). Finally, the current pose is estimated from the latent variable by inverse mapping of pose dimensionality reduction.

8. Experiments

We conducted experiments with general tight-fitting and loose-fitting clothing (shown in Fig. 8). A subject wearing each clothing moved while 8 roof-mounted synchronized cameras captured the subject at 30 fps (1024 $\times$ 768 pixels). Three kinds of motions were selected: dance, exercise, and (simple) walking.Note that our method can learn any kind of clothing and motion. Subjects wore tight-fitting clothing in exercise sequences while they wore loose-fitting clothing in other sequences. The reconstruction voxel size was 10^3mm. For obtaining pose data, IGS-190[21] was used. The subject put it on under clothing. With this MoCap, 54 dimensional pose data (i.e., 18 3-DOF joints) was obtained in synchronization with the images.

In learning processing, a volume was reconstructed at each frame by SFS and mesh deformation[29] and then converted to its volume descriptor, whose dimension was 2×16 (azimuth divisions) $\times 5$ (height divisions) $= 160$-D. For each kind of motion, 300 training frames were used for preparing a volume latent space, whose dimension was empirically determined to be 6, by using GPDM[15]. For comparative experiments, another volume latent space was also generated from the same sample volumes using PCA. The dimension of a pose latent space was determined so that its cumulative percentage was over 0.95: 4-D for all kinds of the motions.

With the motion prior obtained from one subject, each of the five methods below was applied to all test sequences of five subjects, each of which consisted of 300 frames:

Table 1. RMS errors (degrees) of estimated joint angles.

	8 cameras		
	Dance	Exercise	Walking
M1. Direct detection	6.13	10.21	8.52
M2. D with PCA	6.34	12.77	7.45
M3. D with GPDM	5.81	7.69	6.90
M4. T with GPDM	5.39	6.46	5.09
M5. Proposed method	5.04	5.03	3.77
	4 cameras		
M1. Direct detection	13.77	13.81	11.34
M2. D with PCA	14.71	16.7	8.64
M3. D with GPDM	6.73	8.03	8.42
M4. T with GPDM	6.37	6.46	6.27
M5. Proposed method	5.59	5.54	4.91

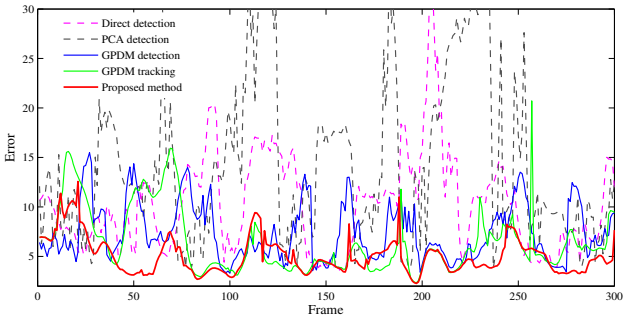
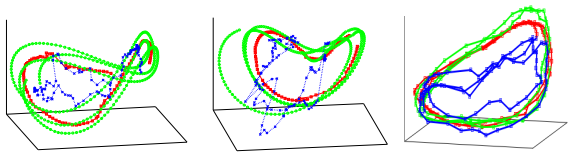


Figure 6. Comparison of joint angle errors in dance sequences.



Dance Exercise Walking

Figure 7. Tracking results in the latent space that is 3D for visualization (green: sample refined volumes, blue: input visual hulls, red: our tracking results). Left: dance1, Right: dance2.

M1. Direct detection The pose is regressed from the input visual hull without via their latent spaces.

M2/M3. Detection with PCA/GPDM The input visual hull was mapped to the volume latent space, generated by PCA/GPDM, and it was used for pose regression.

M4. Tracking with GPDM This was same as the proposed method, except without the visual hull constraint.

M5. Proposed method 256 particles were used. The weight variables for particle likelihood computation were $\nu = 0.07$ and $w = 1$ for a particle and $w = 10$ for an input visual hull.

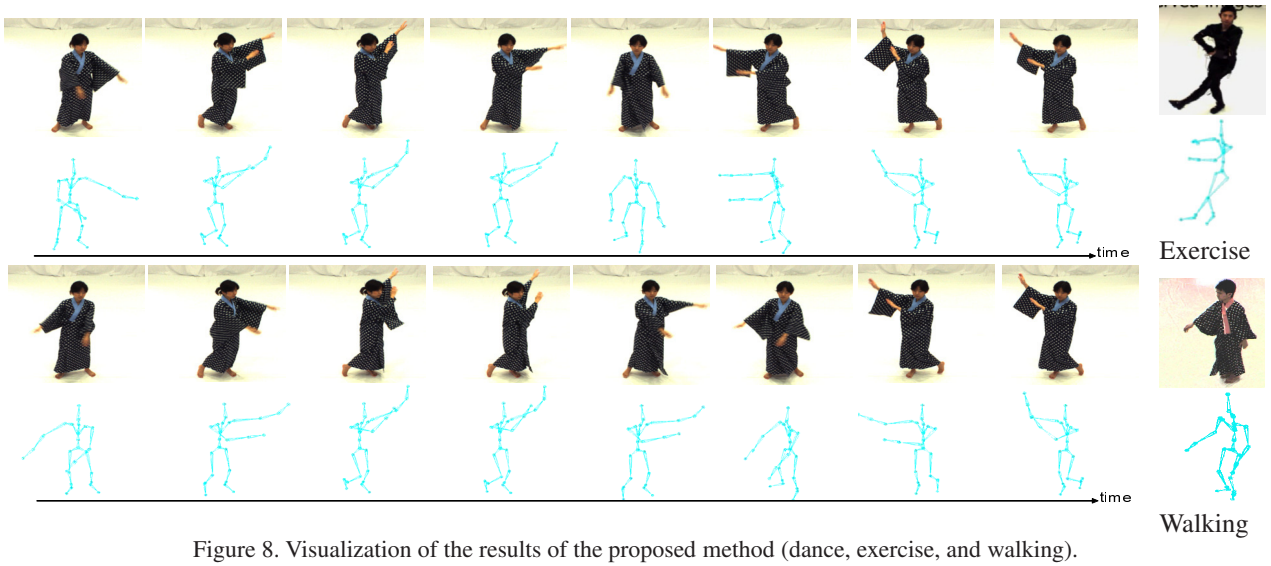


Figure 8. Visualization of the results of the proposed method (dance, exercise, and walking).

Table 1 shows the mean (over all joints, all frames, and all subjects) of the RMS errors of estimated joint angles. The groundtruth was obtained by MoCap. For each sequence, evaluations with 4 and 8 cameras were conducted. Figure 6 shows the temporal histories of the RMS errors of all the joints in a *dance* sequence. From these results, we can make the following observations:

- Nonlinear embedding is superior to linear embedding and without embedding.
- Tracking with motion prior is superior to frame-independent regression.
- The visual hull constraint can improve accuracy, especially in 4 cameras, which produce larger visual hulls than 8 cameras. Errors in exercise sequences, in which tight-fitting clothing was observed, were also suppressed by the constraint. Note that, in these sequences also, the complex poses of a subject made large phantom volumes.

Figure 7 shows the histories of the following three latent variables visualized in a 3D space: sample refined volumes, input visual hulls, and estimated refined volumes acquired by our proposed method. Several examples of joint angles estimated by the proposed method and the method without tracking and the proposed constraints (i.e., M3) are shown in Fig. 8 and 9, respectively.

The proposed method ran at around 3 fps. Although it is much faster (more than 1000 times faster) than existing methods[9, 10] that estimate a human body under clothing, it might not be enough for online applications. The most time-consuming process was the mapping between the volume descriptor and its latent variable. Since this mapping is indispensable for our visual hull constraint, speeding up this mapping is one of the important future challenges (e.g., sparse samples[32]).

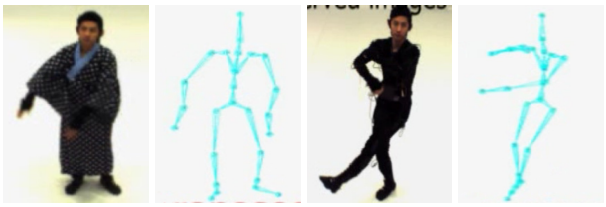


Figure 9. Incorrect pose estimation results by detection with GPDM (i.e., M3). Left: dance (different arm directions), Right: exercise (upside down arms).

9. Concluding Remarks

Our pose estimation method is based on pose regression from the visual hull, which is significantly faster than other pose estimation methods under clothing[9, 10]. The proposed volume descriptor can represent the spatio-temporal variation of the volumes efficiently. The novel visual hull constraint with the volume descriptor can improve matching between the input visual hull and the training refined volumes. This geometric constraint in the observation volume space is integrated with the motion dynamics modeled in the latent space obtained by GPDM. As the result, our method can cope with complex non-rigid shape variations of clothing that hides a human body.

Future work includes improving the visual hull constraint by 1) more precise mapping between the latent spaces and 2) more detailed comparison between the visual hull and the refined volume. One disadvantage of a MoCap using sensors (e.g., accelerometers) is that it cannot implant

the estimated joint angles into the reconstructed volume because the geometric relationship (i.e., relative positions) between the volume and the joint angles measured by sensors is not obtained in the training period. To solve this problem, the geometric relationship can be acquired from the 3D positions of some sensors observable from the cameras. Otherwise, the overlap between the skeleton and the reconstructed volume might be able to be determined based on several characteristic body features (e.g., head, face) both in the training and online processing periods.

The GPDM code and the deformable mesh software were provided by courtesy of Neil Lawrence and Shohei Nobuhara, respectively. We also would like to thank Noriaki Ichida for extensive experiments. Our deepest thanks are extended to Takeo Kanade for useful comments.

References

[1] R. Poppe, "Vision-based human motion analysis: An overview," *CVIU*, Vol.108, No.2, pp.4–18, 2007.

[2] H. Sidenbladh, M. J. Black, and L. Sigal, "Implicit Probabilistic Models of Human Motion for Synthesis and Tracking," *ECCV*, 2002.

[3] A. Agarwal and B. Triggs, "Tracking Articulated Motion using a Mixture of Autoregressive Models," *ECCV*, 2004.

[4] Y. Furukawa and J. Ponce, "Accurate, Dense, and Robust MultiView Stereopsis," *CVPR*, 2007.

[5] G. Cheung, T. Kanade, J. Bouguet, and M. Holler, "A real time system for robust 3D voxel reconstruction of human motions," *CVPR*, 2000.

[6] X. Wu, O. Takizawa, and T. Matsuyama, "Parallel Pipeline Volume Intersection for Real-Time 3D Shape Reconstruction on a PC Cluster," *The 4th IEEE International Conference on Computer Vision Systems (ICVS)*, 2006.

[7] I. Mikic, M. Trivedi, E. Hunter, and P. Cosman, "Human Body Model Acquisition and Tracking using Voxel Data," *IJCV*, Vol.53, No.3, pp.199–223, 2003.

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

[9] B. Rosenhahn, U. Kersting, K. Powell, R. Klette, G. Klette, H.-P. Seidel, "A system for articulated tracking incorporating a cloth model," *Machine Vision and Applications*, Vol.18, No.1, pp.25–40, 2007.

[10] A. O. Balan and M. J. Black, "The Naked Truth: Estimating Body Shape Under Clothing," *ECCV*, 2008.

[11] N. Ukita, R. Tsuji, and M. Kidode, "Real-time Shape Analysis of a Human Body in Clothing using Time-series Part-labeled Volume," *ECCV*, 2008.

[12] N. How, M. Leventon, and W. Freeman, "Bayesian Reconstruction of 3D Human Motion from Single-Camera Video," *NIPS*, 1999.

[13] M. Brand, "Shadow Puppetry," *ICCV*, 1999.

[14] N. D. Lawrence, "Probabilistic non-linear principal component analysis with Gaussian process latent variable models," *Journal of Machine Learning Research*, Vol.6, pp.1783–1816, 2005.

[15] J. M. Wang, D. J. Fleet, A. Hertzmann, "Gaussian Process Dynamical Models for Human Motion," *PAMI*, Vol.30, No.2, pp.283–298, 2008.

[16] N. D. Lawrence, "Local distance preservation in the gp-lvm through back constraints," *International Conference on Machine Learning*, 2006.

[17] C. H. Ek, P. H. S. Torr, and N. D. Lawrence, "Gaussian Process Latent Variable Models for Human Pose Estimation," *4th International Workshop on Machine Learning for Multimodal Interaction*, 2007.

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

[19] J. MacCormick and A. Blake, "A probabilistic exclusion principle for tracking multiple objects"' *ICCV*, 1999.

[20] J. Sullivan, A. Blake, M. Isard, and J. MacCormick, "Bayesian object localization in images," *IJCV*, Vo.44, pp.111–135, 2001.

[21] Xsens Technologies B.B., moven, http://www.moven.com/en/home_moven.php.

[22] A. Agarwal and B. Triggs, "3D Human Pose from Silhouettes by Relevance Vector Regression," *CVPR*, 2004.

[23] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *PAMI*, Vol.24, No.4, pp.509–522, 2002.

[24] Y. Sagawa, M. Shimosaka, T. Mori, and T. Sato, "Fast online human pose estimation via 3D voxel data," *IROS*, 2007.

[25] M. Kortgen, G.-J. Park, M. Novotni, and R. Klein, "3D shape matching with 3D shape contexts," *7th Central European Seminar on Computer Graphics*, 2003.

[26] A. Frome, D. Huber, R. Kolluri, T. Bulow, and J. Malik, "Recognizing Objects in Range Data Using Regional Point Descriptors," *ECCV*, 2004.

[27] B. Bustos, D. A. Keim, D. Saupe, T. Schreck, and D. V. Vranic, "An experimental effectiveness comparison of methods for 3D similarity search," *International Journal on Digital Libraries*, Vol.6, No.1, pp39–54, 2006.

[28] K. N. Kutulakos and S. M. Seitz, "A Theory of Shape by Space Carving," *IJCV*, Vol.38, No.3, pp.199–218, 2000.

[29] S. Nobuhara and T. Matsuyama, "Deformable Mesh Model for Complex Multi-Object 3D Motion Estimation from Multi-Viewpoint Video," *3DPVT*, 2006.

[30] N. Ahmed, C. Theobalt, P. Dobrev, H.-P. Seidel, and S. Thrun, "Robust Fusion of Dynamic Shape and Normal Capture for High-quality Reconstruction of Time-varying Geometry," *CVPR*, 2008.

[31] A. Thayananthanm R. Navaratnam, B. Stenger, P. H. S. Torr, and R. Cipolla, "Multivariate relevance vector machines for tracking," *ECCV*, 2006.

[32] E. Snelson and Z. Ghahramani, "Sparse Gaussian Processes using Pseudo-inputs," *NIPS*, 2005.