Real-Time Shape Analysis of a Human Body in Clothing Using Time-Series Part-Labeled Volumes

Norimichi Ukita, Ryosuke Tsuji, and Masatsugu Kidode

Nara Institute of Science and Technology, Japan
ukita@ieee.org

Abstract. We propose a real-time method for simultaneously refining the reconstructed volume of a human body with loose-fitting clothing and identifying body-parts in it. Time-series volumes, which are acquired by a slow but sophisticated 3D reconstruction algorithm, with body-part labels are obtained offline. The time-series sample volumes are represented by trajectories in the eigenspaces using PCA. An input visual hull reconstructed online is projected into the eigenspace and compared with the trajectories in order to find similar high-precision samples with body-part labels. The hierarchical search taking into account 3D reconstruction errors can achieve robust and fast matching. Experimental results demonstrate that our method can refine the input visual hull including loose-fitting clothing and identify its body-parts in real time.

1 Introduction

Using human motion information, a number of real-world applications can be realized; for example, gesture-based interface, man-machine interaction, and CG animation and computer-supported study of sports/expertises. For acquiring that information, *body-part identification* (i.e., posture estimation) is an essential technique. Such techniques have been proposed in many studies [1].

In a method based on 2D information obtained by a single camera, human posture is estimated by fitting an approximate human-body model into a human region in an image. The estimation result is, however, not robust to occlusions.

To improve the robustness to occlusions, 3D volume reconstruction from multiple views is effective. A reconstructed volume is useful not only for posture estimation but also for 3D shape analysis. Although 3D reconstruction requires a computational cost in general, Shape-From-Silhouette (SFS) can provide the volume (i.e., visual hull) of a moving person stably in real time [2,3]. Online applications using 3D shape and posture are, therefore, feasible by speeding up 3D posture estimation following 3D reconstruction.

In most methods based on a 3D volume, as with those based on a 2D image, the posture is estimated so that the overlapping region between the reconstructed

D. Forsyth, P. Torr, and A. Zisserman (Eds.): ECCV 2008, Part III, LNCS 5304, pp. 681–695, 2008.

volume and the 3D human model that consists of *simple rigid* parts (e.g., cylinders) is maximized (see [4,5], for example). The body parts (e.g., torso and arms) can be given manually or detected from time-series reconstructed volumes by extracting sub-volumes, each of whose motion is regarded as a rigid motion (see [6,7], for example). All of these methods can work well under the assumption that each body part can be approximately modeled as a rigid part. Approximation errors can be reduced by using a precise human-model obtained by a 3D scanner and by estimating the shape deformation around a joint [8]. In [9], the regions of bending limbs can be identified in the reconstructed volume without the assumption of rigidity. Even with these methods, it is impossible to represent *a large variation* of the shape of fully *non-rigid loose-fitting* clothing.

Although the shapes/motions of clothing are modeled and simulated in several applications (e.g., CG [10] and non-rigid tracking [11]), it is impossible to estimate the shapes/motions without information about human motion. In [12], the shapes of a skirt and legs in it are reconstructed simultaneously using a clothing model. Although this method might be most successful, the observed target is simple (i.e., simple deformation without occlusions) and the computational cost is very expensive (5min/frame). As far as we know, no existing algorithm can simultaneously achieve shape reconstruction and posture estimation of a human body with *loose-fitting clothing* in complex motion in *real-time*.

In addition to this essential problem, the previous methods tend to fail due to 3D reconstruction errors. Especially using SFS, *ghost volumes* must be included in concave areas of a human body even if the pre-processes (e.g., camera calibration and silhouette extraction) are achieved without any error. The ghost volumes and other reconstructed errors can be refined based on post-processes such as multi-view photo consistencies (e.g., space curving [13]) and additional restrictions such as silhouette consistencies and temporal smoothness (e.g., deformable mesh model [14]). Similar sophisticated algorithms allow us to fulfill photo consistencies of a textureless object [15] and to deform a visual hull with silhouette consistencies and estimated surface normals using a template human model [16,17]. Even occluded clothes can be reconstructed using color markers printed on a surface and hole filling [18,19]. None of these methods, however, can achieve real-time 3D reconstruction of a surface with arbitrary texture.

Based on the above discussions, we propose a method for analyzing the shape of a human body in loose-fitting clothing, which has the properties below: (1) real-time processable for online applications, (2) identifying body parts with significantly deformable clothing, and (3) refining the volume with arbitrary textures. With our body-part identification, each voxel in the reconstructed volume including clothing is classified into a body-part label. The result does not show the joint positions/angles (i.e., posture) but enables robust posture estimation using existing methods and their extension to posture estimation of a body wearing loose-fitting clothing; each joint must be in the boundary of body parts estimated by our method. The purpose of our volume refinement is to fill and remove significant errors due to the failure in silhouette extraction and SFS.

2 Basic Schemes for Analyzing Loose-Fitting Clothing

In many methods for body-part identification and posture estimation, knowledge about the human body is useful for improving accuracy and robustness. In recent years, several methods learn and employ observed human motions as training samples (e.g., the movable range of each joint angle [20] and the probabilistic representation of each joint motion [21,22]). The example-based approach is superior to a parametric representation in terms of correctly representing complicated and small variations of the posture and motion.

In most of the previous example-based methods, the motion data is expressed by a set of joint angles obtained by using a Motion Capture system with markers. Using the real data obtained by MoCap is superior to using CG samples [23] in terms of reality. For our purposes, however, the following problems arise in using MoCap: (1) when a person wears loose-fitting clothes, the joint positions cannot be measured because the markers on the clothes cannot stay in their corresponding joints, and (2) total shape information cannot be obtained because only the 3D positions of the markers are measured. Even with a number of markers attached on the surface of a target [24], detailed shape analysis is difficult because of interpolation among the markers and large holes caused by occlusions.

To realize our objectives while keeping the advantages of the example-based methods (i.e., reality), therefore, we employ the following training samples:

- The time-series reliable volumes of a human body wearing clothing; the reliable volumes are reconstructed with less errors by employing a slow but sophisticated method such as [13,14].
- The body-part labels of each voxel in the total shape (i.e., volume).

These allow us to have the following advantages: (1) learning without any marker that prevents free motions of a body and clothing and (2) analyzing not sparse points on the surface of a body but its volume.

In our training scheme, the body-part label in each voxel is obtained from a reconstructed sample volume wearing clothes in which each body part is colored with a different color. In online analysis, the sample volumes with the labels are compared with an input volume (i.e., visual hull) reconstructed online in order to find similar samples. Using PCA, all the volumes are analyzed in a lower-dimension eigenspace for quick search; (1) the input visual hull is projected into the eigenspace in order to find samples similar to it, and then (2) the reliable part-labels are acquired from the samples. Although a distinctive 3D shape descriptor (e.g., [25]) is effective for similarity retrieval, it needs more computational cost. In this work, therefore, characteristic features are extracted from the reconstructed volume with PCA for real-time search.

In this paper, one of the following 10 part-labels is allocated to each voxel in a human body; *head*, *torso*, *right upper-arm*, *right forearm*, *left upper-arm*, *left forearm*, *right thigh*, *right lower-leg*, *left thigh*, and *left lower-leg* labels. In addition, a special label *non-object* is prepared in order to be allocated to ghost volumes. By allocating these 11 labels to the input visual hull, body-part labeling and volume refinement are realized simultaneously.

Time-series sample volumes of each sequence are represented by a trajectory in the eigenspace as with a manifold in the parametric eigenspace method [26]. In terms of dealing with 3D volumes, our problem has the following distinctive difficulties:

Huge dimensions. The voxels in a human volume is numerous. Since dimension reduction using PCA is insufficient for real-time processing, dual hierarchical searches in the eigenspace are implemented (Sec 3.3 and 4.3).

Difference between an input visual hull and samples. While a sample volume is refined, an input visual hull may include large amounts of ghost volumes. A matching scheme robust to this difference is required (Sec. 4.2).

3 Time-Series Volume Learning

3.1 Generating Reliable Volumes with Part-Labels

The visual hull of a human body with part-colored clothing (Figure 1 (a)) is reconstructed by SFS. Ghost volumes and other errors are then refined using the deformable mesh model [14] as shown in Figure 1 (b) in order to approximate the real volume for preparing samples. Next, the part-labeled image (Figure 1 (c)) is generated by color detection. The colors of the part-labeled images are projected onto the refined volume from multiple viewpoints in order to allocate one of the part-labels to each surface voxel. The inside voxels are labeled by finding the nearest surface voxel. Finally, the reliable volume with the part-labels can be acquired as shown in Figure 1 (d).

3.2 Volume Learning Based on PCA

For PCA, the dimensions of the volume (i.e., the number of voxels) in all frames must be identical. Therefore, the voxels in a fixed-size 3D bounding box, which is defined so that its centroid coincides with that of the volume in each frame, are extracted. The size of the bounding box is determined so that it can cover the whole-body in respective frames.

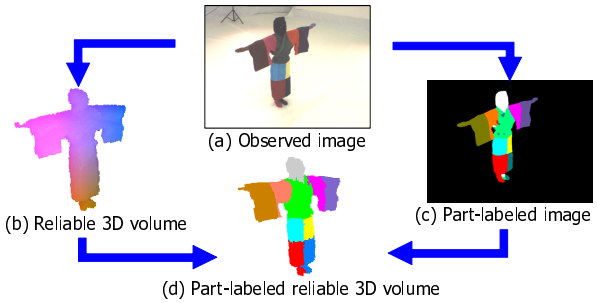


Fig. 1. Process flow for generating reliable volumes with part-labels.

Let $\boldsymbol{v}_t = (v_{t,1}, v_{t,2}, \cdots, v_{t,d})^T$ ($v_{t,i} \in \{0, 1\}$, where 0 and 1 denote non-object and body voxels, respectively) be d-dimensional voxels observed at time t (i.e., Fig. 1 (b)). If T sample volumes are observed in total, a matrix consisting of the sample volumes is expressed by $V = (\boldsymbol{v}_1 - \boldsymbol{m}, \boldsymbol{v}_2 - \boldsymbol{m}, \cdots, \boldsymbol{v}_T - \boldsymbol{m})$, where $\boldsymbol{m}$ denotes the average of T volumes. The covariance matrix of the sample volumes, $S = VV^T$, is computed in order to acquire a set of eigenvectors, $\{\boldsymbol{e}_i | i \in \{1, \cdots, d\}\}$ (in ascending order), of S. With the first k ($<< d$) eigenvectors, a d-dimensional volume $\boldsymbol{v}_t$ can be approximated by a smaller dimensional vector as follows: $\boldsymbol{v}_t$ is transformed to $\boldsymbol{y}_t$ in the k-dimensional eigenspace by the following linear projection using a matrix $E = (\boldsymbol{e}_1, \cdots, \boldsymbol{e}_k)$ consisting of the k eigenvectors:

$$\boldsymbol{y}_t = E^T (\boldsymbol{v}_t - \boldsymbol{m}) \tag{1}$$

By concatenating the projected volumes of each observed sequence in order of observed time, time-series volume variations can be represented by a trajectory.

Let $\{\boldsymbol{y}_1^L, \cdots, \boldsymbol{y}_T^L\}$ be samples projected into the eigenspace. Each projected point in a trajectory has its original volume with the part-labels (i.e., Fig. 1 (d)). Figure 2 shows examples of the trajectory and the volumes with the part-labels. From the figure, it can be confirmed that similar volumes are located nearby.

3.3 Hierarchical Volume Learning

The detailed shape of a target is lost in significantly lowered dimensions. In our method, for real-time search while retaining the detailed shape, the volume is analyzed in two stages, namely low-resolution whole-body and high-resolution body-part analyses. After identifying the rough locations of body parts in the whole-body analysis, the detailed shape is acquired in the high-resolution analysis of each body-part. For the hierarchical analysis, the processes mentioned in Sec. 3.2 (i.e., division into bounding boxes and dimension decreasing based on PCA) are applied not only to a whole-body volume but also to all high-resolution body parts. Consequently, 11 eigenspaces (i.e., a whole-body + 10 body parts) are generated in total. Examples of the bounding boxes are shown in Figure 3. Note that a bounding box may partially overlap other bounding boxes.

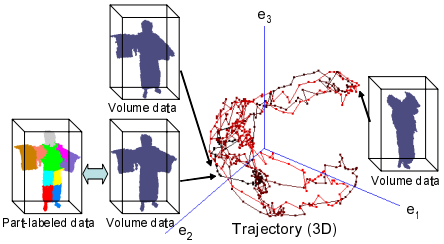
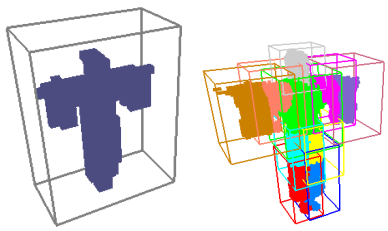


Fig. 2. Trajectory with part-labels in a (3D) eigenspace



(a) Whole body (b) Body parts

Fig. 3. Bounding boxes

4 Shape Analysis: Part-Labeling and Volume Refinement

Figure 4 illustrates the process flow of online shape analysis. Each number in the figure indicates the section that introduces the corresponding process in detail.

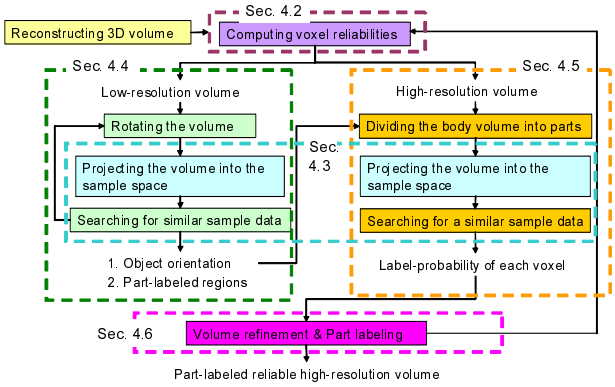


Fig. 4. Process flow of shape analysis

4.1 Search from Time-Series Sample Volumes

In this Section 4.1, the outline of a search algorithm that is employed both for whole-body and body-parts analyses is described.

Time-series volumes at the current frame and n previous frames are projected into the eigenspace using Formula (1). Let $\boldsymbol{Y}_t^I = (\boldsymbol{y}_{t-n}^I, \boldsymbol{y}_{t-n+1}^I, \cdots, \boldsymbol{y}_t^I)$, where $\boldsymbol{y}_t^I$ denotes the projected point of the input visual hull at t, be the input trajectory in the eigenspace. By comparing the input trajectory $\boldsymbol{Y}_t^I$ and sub-trajectories $\boldsymbol{Y}_s^L = (\boldsymbol{y}_{s-n}^L, \boldsymbol{y}_{s-n+1}^L, \cdots, \boldsymbol{y}_s^L)$ (where $s \in \{n+1, \cdots, T\}$) of the trajectory consisting of sample volumes, sample trajectories similar to the input one are searched for. In general, as the number of the previous frames (i.e., n) increases, the stability of the search improves. The increase in n, on the other hand, causes the problems below:

- The search speed gets slow.
- Since a sample sequence must coincide with the input sequence for a long time, a partially-matched sample sequence (i.e., a sequence similar to the input sequence for a short time) is ignored.

Therefore, n should be determined in accordance not only with search stability but also with processing time and applicability to recognizing the combination of the short-period motions, depending on an application; $n = 5$ in our experiments.

For search in the eigenspace, the summation of the distances between corresponding points of the input and sample sequences is evaluated as the similarity

between them. Practically, $\boldsymbol{y}_s^L$ in $\boldsymbol{Y}_s^L$ that leads the following minimum summation D_t is considered to be most similar to the input $\boldsymbol{y}_t^I$:

$$D_t = \min_{s=n+1}^{T} \sum_{i=-n}^{0} \| \boldsymbol{y}_{s+i}^L - \boldsymbol{y}_{t+i}^I \|, \tag{2}$$

where $\| \cdot \|$ denotes the norm of a vector.

If a selected sample shows the shape/posture of an online observed person correctly, the reliable volume with part-labels can be obtained from the selected sample. It is, however, difficult to find such a sample because the input visual hull may include many errors while all the sample volumes are refined to be without errors. Furthermore, the processing time increases in proportion to the number of samples if all the samples are compared with the input visual hull. To cope with these problems, the following is achieved in our method:

Voxel Reliability. The occurrence probability of the ghost volume in each voxel is evaluated. We call this probability *voxel reliability*.

Efficient Search. The nearest neighbor of the input visual hull is found from a small number of samples based on hierarchical groups of all samples.

In what follows, these solutions for the ghost volumes and the processing time are introduced in Sections 4.2 and 4.3, respectively, and then the procedures of our online shape analysis is described in practical order.

4.2 Ghost Volume Suppression Using the Voxel Reliability in Search

It is hard to predict from observed images where 3D reconstruction errors due to the failure in silhouette extraction occur. This is because the difficulty in silhouette extraction changes depending not only on a target but also on a background scene and image noises. On the other hand, ghost volumes generated due to SFS occur in concave areas of a target. That is, the voxel reliability can be analyzed only in accordance with the shape of the target.

The voxel reliability is given to each voxel in all sample volumes. The voxel reliabilities in each refined sample volume can be estimated by comparing it with "its visual hulls with ghost volumes reconstructed by SFS". Note that the ghost volumes change depending not only on the shape of the volume but also on the geometric configuration among the target and the cameras. Although it is very troublesome and difficult to obtain the visual hulls in various positions/orientations of the target in real observations, the visual hulls can be virtually produced from the refined sample volume as follows:

1. Assume that the extrinsic parameters of real cameras have achieved. The projective silhouette of a sample volume, which is virtually located in visual fields of the cameras, is obtained in each camera.
2. The visual hull is computed from the obtained silhouettes using SFS. In each voxel, the difference between the sample volume and the computed visual hull means a ghost volume.

3. Steps 1 and 2 are executed while changing the location/rotation of the sample volume in order to obtain a number of the visual hulls.
4. In each voxel of the visual hulls, the number of the ghost volumes is counted and its rate (denoted by $r_{t,v}$) is computed by $r_{t,v} = \frac{1}{N_{var}} \sum_{i}^{N_{var}} \delta_{t,v}^{i}$, where
   - N_{var} denotes the number of observations in the virtual 3D space and
   - $\delta_{t,v}^{i}$ outputs 0 if the reconstructed results of v-th voxel in t-th volume are different between the original reliable volume and i-th positions/orientations in the virtual 3D space, otherwise 1.

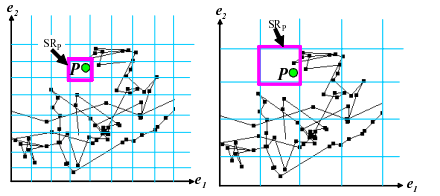
   $r_{t,v}$ is the voxel reliability of v-th voxel in t-th sample volume.
5. The voxel reliability corresponding to the reliable t-th volume (denoted by R_t) is expressed by the following matrix: $R_t = \mathrm{diag}\,(r_{t,1}, \cdots, r_{t,d})$.
6. 1, 2, 3, 4, and 5 are executed for every sample volume for acquiring the voxel reliabilities in all sample volume.

With the following projection formula given by applying the voxel reliability to Formula 1, the projected point of an input visual hull, v_t^I, observed at t (denoted by $\hat{y}_t^I$) is obtained: $\hat{\boldsymbol{y}}_t^I = E^T R_t \left(\boldsymbol{v}_t^I - \boldsymbol{m}\right)$. Using $\hat{\boldsymbol{y}}_t^I$, adverse effects of ghost volumes in search can be reduced. To estimate the voxel reliability R_t, however, the sample corresponding to the input visual hull $\boldsymbol{v}_t^I$ is required. This is the chicken-and-egg problem. Under the assumption that the difference between subsequent volumes is very small, the voxel reliability at t is estimated based on the sample selected at $t-1$.

4.3 Efficient Search from Time-Series Sample Volumes

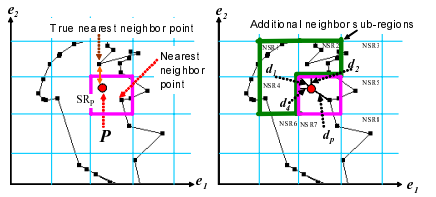
For speeding up search in the eigenspace, the likelihood expressed by Formula 2 should be evaluated for a small number of samples near an input projected point. To determine the limited samples, therefore, the eigenspace is divided into several sub-regions and the samples in each sub-region is counted in advance.

In online search, the sub-region SR_P including the input projected point P is first selected. The similar samples are then found from the samples in SR_P. Figure 5 illustrates examples. As a sub-region becomes smaller as shown in Figure 5 (a), the number of selected samples decreases. However, SR_P may have no sample as shown in (a). To cope with this problem, larger sub-regions are searched following a small sub-region (shown in Figure 5 (b)).



(a) m-divided space (b) n-divided space ($n < m$)

Fig. 5. Efficient hierarchical search



(a) Search failure (b) Search-regions expansion

Fig. 6. Additional search regions

However, the true nearest neighbor (indicated by "True nearest neighbor point" in Figure 6 (a)) may exist in a sub-region other than SR_P. Let (1) NSR_i be one of the sub-regions neighboring SR_P and (2) d_i and d_p be the distances from P to NSR_i and the nearest neighbor in SR_P, respectively, as illustrated in Figure 6 (b). To always find the true nearest neighbor, the neighboring sub-regions, each of which satisfies $d_i < d_p$, are also searched.

With the above mentioned search algorithm, both high-speed capability and search stability for always finding the nearest neighbor can be attained.

4.4 Rough Shape Analysis of the Low-Resolution Whole-Body

In what follows, the procedures using the above mentioned functions are described in practical order.

A low-resolution volume reconstructed online is projected into the eigenspace in order to estimate the following two information:

Target orientation. Each volume is reconstructed in the world coordinate system. That is, even if the shapes/postures of a person are identical, the reconstructed volumes of the different orientations are different. The orientations of the volumes must be, therefore, aligned.

Body-parts regions. In order to compare the samples of each high-resolution body-part with the region of this body-part in an input visual hull, this region must be identified in the input.

This information is estimated as follows:

1. The centroid of the input visual hull is estimated.
2. The input visual hull is rotated θ along the vertical axis through the centroid.
3. The volume inside the bounding box, whose size is identical to the size of the whole-body in sample training, is projected into the eigenspace of the whole-body. n previous volumes are also projected.
4. With D_t in Formula (2), the sample nearest the projected point is selected.
5. Steps 2, 3, and 4 are repeated while changing θ.
6. The orientation of the input visual hull (denoted by $\acute{\theta}$) is determined so that D_t with regard to $\acute{\theta}$ is minimum. The input visual hull is then rotated $\acute{\theta}$ and the sample volume corresponding to the minimum D_t is selected as the nearest neighbor.
7. The bounding box including each body part in the input visual hull is acquired by overlapping the selected sample volume with its body-part labels and their bounding boxes.

4.5 Detailed Shape Analysis of the High-Resolution Body-Parts

The volume inside the bounding box of each body part is then analyzed in order to estimate the *label probabilities* in the high-resolution volume. The label probabilities in each voxel mean how appropriate 11 labels, including *non-object*, are to the label of this voxel.

1. An input high-resolution visual hull is rotated $\acute{\theta}$ and then divided into the regions of its body parts.
2. The volume inside each bounding box is projected into the corresponding eigenspace. n previous volumes are also projected. The following steps are executed for each body-part.
3. D_t in Formula (2) is evaluated. Samples, each of whose D_t is less than a predefined threshold, are selected.
4. Let N_{data} be the number of the selected samples. The likelihood of the selected sample d (denoted by L_t^d, where $d \in \{1, \cdots, N_{data}\}$) is the inverse of D_t of d. With this likelihood, the probability of l-th part label in v-th voxel (denoted by $P^{v,l}$) is defined by $P^{v,l} = \sum_{d=1}^{N_{data}} \frac{L_t^d}{S_t} \epsilon_{v,l}^d$, where $S_t = \sum_{d=1}^{N_{data}} L_t^d$ and $\epsilon_{v,l}^d$ outputs 1 if the part-label of v-th voxel in d is l, otherwise 0.

4.6 Part-Labeling and Volume Refinement

By integrating the label probabilities estimated in all bounding boxes, the high-resolution whole-body volume with the label probabilities is generated. Note that several bounding boxes overlap with each other. In such an overlapping region, the label probabilities in the same voxel may be different among different bounding boxes. The average of the label probabilities with regard to each body-part label (i.e., $\frac{1}{N_{part}} \sum_{p=1}^{N_{part}} P^{p,v,l}$, where N_{part} denotes the number of physical body parts, namely 10 in our method, and $P^{p,v,l}$ denotes $P^{v,l}$ in body-part p) is given to a voxel. In addition, spatially neighboring voxels should have the same label because each body part must be one cluster. For harmonization among the neighboring voxels, therefore, the weighted average of the label probabilities in $N_{nv} \times N_{nv} \times N_{nv}$ neighboring voxels (denoted by$\hat{P}^{v,l}$) is then computed in each voxel: $\hat{P}^{v,l} = \frac{1}{\sum_{i=1}^{N_{nv}^3} w_i} \sum_{i=1}^{N_{nv}^3} w_i \bar{P}^{i,l}$, where w_i denotes the distance between v and its neighboring voxel i. In our experiments, $N_{nv} = 3$. Finally, in each voxel in the whole body, the label with the maximum label probability $\hat{P}^{v,l}$ is considered to be the body-part label.

5 Experiments

We conducted experiments with loose-fitting 10-colored clothing as shown in Figure 1 (a). A person wearing this clothing danced while seven synchronized cameras captured this person at 30 fps (1024 × 768 pixels).

5000 frames were captured for training samples. All the sample images were prepared with one subject. With these images, the time-series volumes of the whole-body were reconstructed. The voxel sizes in low and high resolution volumes were 60mm^3 and 20mm^3, respectively. While the voxel dimension of a low-resolution volume was always $23 \times 11 \times 31 = 8743$, those of high-resolution volumes were changed depending on the body part (between 7479 and 26825). The dimension number of each eigenspace is determined based on the cumulative contribution rate. In our experiments, the dimension number, which allows

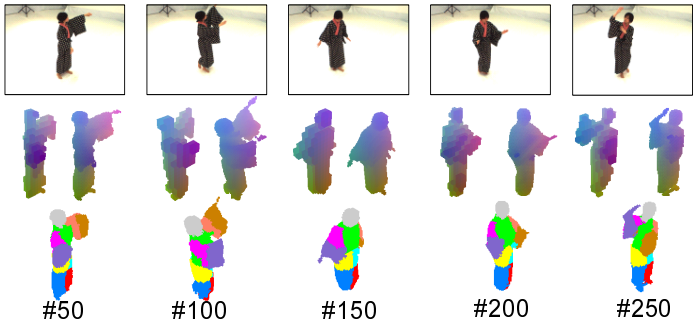


Fig. 7. Experimental results (1st-row: Observed images, 2nd-row: Input low(left)/high(right)-resolution visual hulls, 3rd-row: Shape-analysis results).

us to analyze the human volumes with satisfactory accuracy, was empirically determined so that the cumulative contribution rate was 75%, which is obtained using 44 eigenvectors in the whole-body volumes, for example.

Using these samples recorded in the eigenspaces, shape analysis was conducted. For input image sequences, two subjects were observed separately; one (155 cm) was the person who was captured for the samples. Since their heights were different (155cm and 175cm), the reconstructed volumes of the other were normalized based on the ratio between their heights. They wore the same-shape clothing as that used for the samples, and performed the same dance as the samples. Partial results were shown in Figure 7. An example of shape refinement is shown in Figure 8. This result demonstrates that ghost volumes generated by being surrounded with two arms could be removed. An example of the results for the other subject is shown in Figure 9. It can be confirmed that the volume could be refined even if the input visual hull was corrupted due to the failure of silhouette extraction as indicated by red circles.

We also conducted experiments using tight clothes under the same condition as that of the above experiments; samples with the tight clothes were prepared. An example of the results is shown in Figure 10. It can be confirmed that our method is also effective for this kind of common clothing.

The processing time using a PC with an Opteron 1.8GHz was 24 fps. Although this is a little slower than the video rate (30fps), a moving person can be observed as shown in the results.

As explained above, we can confirm that our proposed method almost satisfies our purposes introduced at the end of Section 1, namely real-time processing, body-part identification, and volume refinement.

For quantitative evaluation, the following two volumes that were acquired by observing colored clothing used for the samples were compared:

True volume. Volumes reconstructed by the method for obtaining samples by using color information.

Result. Volumes acquired without color information by the proposed method.

Table 1. Error analysis: The percentages of false-positive (FP) and false-negative (FN) voxels in proportion to the size of each body-part are shown

Part	Same subject		Another subject		Tight clothing	
	FP(%)	FN(%)	FP(%)	FN(%)	FP(%)	FN(%)
head	7.2	5.2	7.9	5.7	4.7	5.8
torso	3.9	3.1	5.2	5.8	3.2	3.2
right upper-arm	9.9	6.4	9.1	8.1	4.9	5.6
right forearm	11.4	10.8	13.2	12.2	7.2	6.0
left upper-arm	6.5	11.1	10.1	9.6	4.0	3.6
left forearm	9.5	10.0	14.8	11.0	8.3	7.1
right thigh	5.1	7.1	6.4	4.4	4.8	5.1
right lower-leg	11.8	8.7	11.6	9.4	9.5	7.6
left thigh	3.3	6.3	7.7	5.1	6.8	6.5
left lower-leg	7.5	6.4	12.9	8.1	11.4	9.7

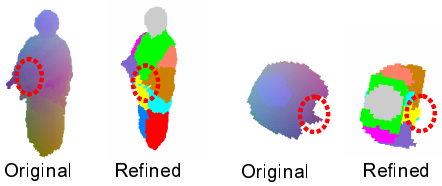


Fig. 8. Volume refinement results shown from two viewpoints

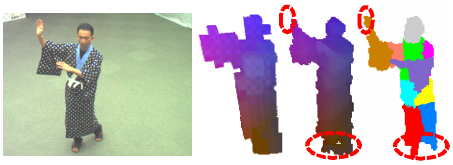


Fig. 9. Results in the other subject

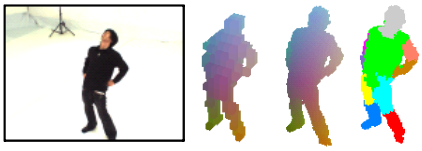


Fig. 10. Results in another dance

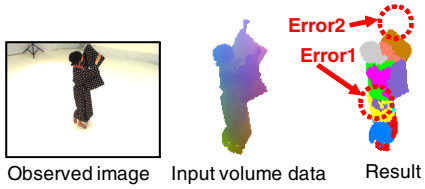


Fig. 11. Wrong labeling results

Table 1 shows the results; errors were increased due to the following factors:

- Subject change between training and analysis; errors in "Another subject" were worse than those in "Same subject".
- Non-rigid motion of loose-fitting clothing; (1) errors in "Tight clothing" were smaller than those in other two results and (2) those in rigid body parts (e.g., head) were smaller than those in non-rigid parts (e.g., forearm).

Although shape analysis was almost successful in most frames, large errors in body-part labeling were caused in a few frames. In an example shown in Figure 11 (Error1), the *right upper-arm* label is allocated to a part of the right thigh.

In analyzing high-resolution body parts, similar samples are searched for independently in each body part. The independent search allows us to recognize a variety of the shapes/postures of the whole-body, each of which is composed by a combination of samples of the body-parts, even if those shapes/postures of the whole body are not included in the samples. However, by simply integrating (i.e., averaging) all the body parts after the independent search, a combination error such as an example in Figure 11 may occur. The consistent results should be acquired by mutually exchanging the search results of the body parts, Furthermore, although our method strongly relies on search for similar samples, useful restrictions about the shape/posture of a human body can improve the method (e.g., "each body must be a cluster" and "the cubic content of each body part should be almost constant").

While a false-negative region in an input visual hull could be recovered in the example shown in Figure 9, a true-positive region was sometimes removed as shown in the example in Figure 11 (Error2). Such an error is caused due to small differences between the input visual hull and samples. These errors are inevitable when new people not included in the samples are observed. To cope with this problem, a number of samples of the same motion should be used. For using a large data set, powerful embedding [27] is useful because it can possibly represent observed volumes well by probabilistic lower-resolution data that has a beneficial effect on discrimination. Its capability also possibly can improve the versatility of the method, which is one of critical limitations of our method; how to generalize to new clothes and new motions. In principle, to recognize them, corresponding volumes have to be in samples. This results in increasing the samples. Therefore, powerful embedding [27] is crucial for the versatility.

6 Concluding Remarks

We proposed an online method for simultaneously refining the reconstructed volume of a human body wearing loose-fitting clothing and identifying body parts in it. In our method, reliable and high-precision time-series volumes with body-part labels are learned in advance. Each volume reconstructed online is compared with these reliable volumes in order to find similar reliable data with body-part labels. The voxel reliability plays a crucial role for matching robust to ghost volumes. Applying PCA to the volumes and hierarchical and coarse-to-fine analyses allow us to speed up and stabilize the search procedure.

This study was supported by National Project on Development of High Fidelity Digitization Software for Large-Scale and Intangible Cultural Assets.

References

1. Poppe, R.: Vision-based human motion analysis: An overview. CVIU 108(2), 4–18 (2007)
2. Cheung, G., Kanade, T., Bouguet, J., Holler, M.: A real time system for robust 3D voxel reconstruction of human motions. CVPR 2, 714–720 (2000)

3. Wu, X., Takizawa, O., Matsuyama, T.: Parallel Pipeline Volume Intersection for Real-Time 3D Shape Reconstruction on a PC Cluster. In: The 4th IEEE International Conference on Computer Vision Systems (ICVS) (2006)
4. Mikic, I., Trivedi, M., Hunter, E., Cosman, P.: Human Body Model Acquisition and Tracking using Voxel Data. IJCV 53(3), 199–223 (2003)
5. Hou, S., Galata, A., Caillette, F., Thacker, N., Bromiley, P.: Real-time Body Tracking Using a Gaussian Process Latent Variable Model. In: ICCV (2007)
6. Kakadiaris, I.A., Metaxas, D.: 3D Human Body Model Acquisition from Multiple Views. IJCV 30(3), 191–218 (1998)
7. Cheung, G., Baker, S., Kanade, T.: Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. In: CVPR, vol. 1, pp. 77–84 (2003)
8. Balan, A.O., Sigal, L., Black, M.J., Davis, J., Haussecker, H.W.: Detailed Human Shape and Pose from Images. In: CVPR (2007)
9. Sundaresan, A., Chellappa, R.: Segmentation and Probabilistic Registration of Articulated Body Models. In: ICPR (2006)
10. Bhat, K.S., Twigg, C.D., Hodgins, J.K., Khosla, P.K., Popovic, Z., Seitz, S.M.: Estimating Cloth Simulation Parameters from Video. In: Eurographics/SIGGRAPH SCA (2003)
11. Salzmann, M., Urtasun, R., Fua, P.: Local Deformation Models for Monocular 3D Shape Recovery. In: CVPR (2008)
12. Rosenhahn, B., Kersting, U., Powell, K., Klette, R., Klette, G., Seidel, H.-P.: A system for articulated tracking incorporating a cloth model. Machine Vision and Applications 18(1), 25–40 (2007)
13. Kutulakos, K.N., Seitz, S.M.: A Theory of Shape by Space Carving. IJCV 38(3), 199–218 (2000)
14. Nobuhara, S., Matsuyama, T.: Deformable Mesh Model for Complex Multi-Object 3D Motion Estimation from Multi-Viewpoint Video. In: 3DPVT (2006)
15. Bradley, D., Boubekeur, T., Heidrich, W.: Accurate Multi-View Reconstruction Using Robust Binocular Stereo and Surface Matching. In: CVPR (2008)
16. Vlasic, D., Baran, I., Matusik, W., Popovic, J.: Articulated Mesh Animation from Multi-view Silhouettes. In: ACM SIGGRAPH (2008)
17. Ahmed, N., Theobalt, C., Dobrev, P., Seidel, H.-P., Thrun, S.: Robust Fusion of Dynamic Shape and Normal Capture for High-quality Reconstruction of Time-varying Geometry. In: CVPR (2008)
18. Scholz, V., et al.: Garment Motion Capture Using Color-Coded Patterns. Computer Graphics Forum 24(3), 439–448 (2005)
19. White, R., Crane, K., Forsyth, D.: Capturing and Animating Occluded Cloth. ACM TOG (SIGGRAPH) 26(3) (2007)
20. Herda, L., Urtasun, R., Fua, P.: Hierarchical implicit surface joint limits for human body tracking. CVIU 99(2), 189–209 (2005)
21. Sidenbladh, H., Black, M.J., Sigal, L.: Implicit Probabilistic Models of Human Motion for Synthesis and Tracking. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2350, pp. 784–800. Springer, Heidelberg (2002)
22. Agarwal, A., Triggs, B.: Tracking Articulated Motion using a Mixture of Autoregressive Models. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3023, pp. 54–65. Springer, Heidelberg (2004)
23. Grauman, K., Shakhnarovich, G., Darrell, T.: Inferring 3D Structure with a Statistical Image-Based Shape Model. In: ICCV, pp. 641–648 (2003)

24. Park, S.I., Hodgins, J.K.: Capturing and Animating Skin Deformations in Human Motion. ACM TOG (SIGGRAPH) 25(3), 881–889 (2006)
25. Kazhdan, M., Funkhouser, T., Rusinkiewicz, S.: Rotation invariant spherical harmonic representation of 3D shape descriptors. In: Eurographics/SIGGRAPH SGP, pp. 156–164 (2003)
26. Murase, H., Nayar, S.K.: Visual learning and recognition of 3-D objects from appearance. IJCV 14(1), 5–24 (1995)
27. Lawrence, N.D.: Probabilistic non-linear principal component analysis with Gaussian process latent variable models. Journal of Machine Learning Research 6, 1783–1816 (2005)