Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2007, Article ID 59535, 11 pages
doi:10.1155/2007/59535

Research Article

Motion Segmentation and Retrieval for 3D Video Based on Modified Shape Distribution

Toshihiko Yamasaki and Kiyoharu Aizawa

Department of Information and Communication Engineering, Graduate School of Information Science and Technology, The University of Tokyo, Engineering Building No. 2, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

Received 31 January 2006; Accepted 14 October 2006

Recommended by Tsuhan Chen

A similar motion search and retrieval system for 3D video are presented based on a modified shape distribution algorithm. 3D video is a sequence of 3D models made for a real-world object. In the present work, three fundamental functions for efficient retrieval have been developed: feature extraction, motion segmentation, and similarity evaluation. Stable-shape feature representation of 3D models has been realized by a modified shape distribution algorithm. Motion segmentation has been conducted by analyzing the degree of motion using the extracted feature vectors. Then, similar motion retrieval has been achieved employing the dynamic programming algorithm in the feature vector space. The experimental results using 3D video sequences of dances have demonstrated very promising results for motion segmentation and retrieval.

1. INTRODUCTION

Dynamic three-dimensional (3D) modeling of real-world objects using multiple cameras has been an active research area in recent years [1–5]. Since such sequential 3D models, which we call 3D video, are generated employing a lot of cameras and represented as 3D polygon mesh, realistic representation of dynamic 3D objects is obtained. Namely, the objects' appearance such as shape and color and their temporal change are captured in 3D video. Therefore, they are different from conventional 3D computer graphics and 3D motion capture data. Similar to 2D video, 3D video consists of consecutive sequences of 3D models (frames). Each frame contains three kinds of data such as coordinates of vertices, connection, and color.

So far, researches of 3D video have been mainly focused on its acquisition methods, and they are in their infancy. Therefore, most of the research topics in 3D video were capture systems [1–5] and compression [6, 7]. As the amount of 3D video data increases, the development of efficient and effective segmentation and retrieval systems is being desired for managing the database.

Related works can be found in so-called 3D "motion capture" data aiming at motion segmentation [8–12] and retrieval [13–15]. This is because structural features such as motion of joints and other feature points are easily located and tracked in motion capture data.

For motion segmentation, Shiratori et al. analyzed local minima in motion [8]. The idea of searching local minima in kinematic parameters was also employed in [9]. Some other approaches were proposed based on motion estimation error using singular value decomposition (SVD) [10] and least square fitting [11]. In addition, model-based approaches were also reported using hidden Markov model (HMM) [12] and Gaussian mixture model (GMM) [10].

Regarding content-based retrieval for motion capture data, the main target of previous works [13–15] was fast and effective processing because accurate feature localization and tracking was already taken for granted as discussed above. For instance, an image-based user interface using a self-organizing map was developed in [13]. In [14], motion data of the entire skeleton were decomposed as the direct sum of individual to reduce the dimension of the feature space. Reference [15] proposed qualitative and geometric features opposed to quantitative and numerical features used in previous approaches to avoid dynamic time warping matching, which is computationally expensive.

In contrast to motion segmentation and retrieval for 3D motion capture data, those for 3D video are much more challenging. In motion capture systems, users wear a special suit

with optical or magnetic markers. On the other hand, feature tracking is difficult for 3D video because neither markers nor sensors are attached to the users. In addition, each frame of 3D video is generated independently regardless of its neighboring frames [1–5] due to the nonrigid nature of human body and clothes. This results in unregularized number of vertices and topology, making the tracking problem more difficult.

Therefore, the number of 3D video segmentation algorithms reported so far is quite limited [16–18]. In [16], a histogram of distance among vertices on 3D mesh model and three fixed reference points were generated for each frame, and segmentation was done when the distance between histograms of successive frames crossed threshold values. And, more efficient histogram generation method based on spherical coordinate system was developed in [17]. The problem in these two approaches is that they strongly relied on "suitable" thresholding, which was defined only by empirical study (try and error) for each sequence. In [16, 17], proper threshold setting was left unsolved.

With regard to 3D video retrieval, there are no related works yet except for the one we have developed [19]. However, the development of efficient tools for exploiting a large-scale database of 3D video would become a very important issue in the near future.

The purpose of this work is to develop a motion segmentation and retrieval system for 3D video of dances based on our previous works [18, 19]. To the best of our knowledge, this work is the first contribution to such a problem. We have developed three key components such as feature extraction, motion segmentation, and similarity evaluation among 3D video clips.

In particular, proper shape feature extraction from each 3D video frame and analysis of its temporal change are extra important tasks as compared to motion capture data segmentation and retrieval. Therefore, we have introduced a modified shape distribution algorithm we have developed in [18] to stably extract shape features from 3D models.

Segmentation is an important preprocessing to divide the whole 3D video data into small but meaningful and manageable clips. The segmented clips are handled as minimum units for computational efficiency. Then, a segmentation technique based on motion has been developed [18]. Because motion speed and direction of feature points are difficult to track, the degree of motion is calculated in the feature vector space of the modified shape distribution. The segmentation is achieved by searching local minima in the degree of motion accompanied with a simple verification process.

In retrieving, an example of 3D video clip is given to the system as a query. After extracting the feature vectors from the query data, the similarity to each candidate clip is computed employing dynamic programming (DP) matching [20, 21].

In our experiments, five 3D video sequences of three different kinds of dances were utilized. In the experiments of segmentation, high-accuracy precision and recall rates of 92% and 87%, respectively, have been achieved. In addition, the system has also demonstrated very encouraging results by retrieving a large portion of the desired and related clips.

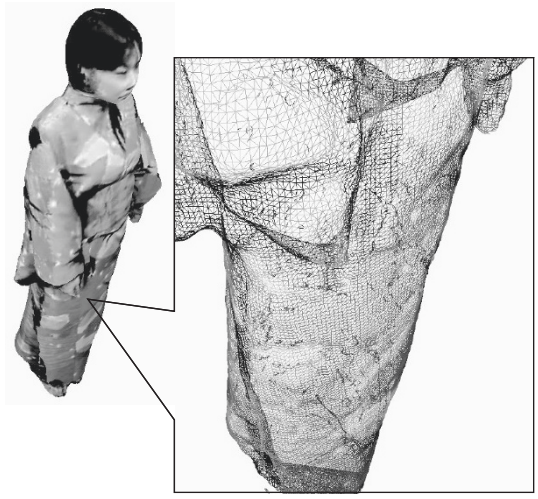


Figure 1: Example frame of our 3D video data. Each frame is described in a VRML format and consists of coordinates of vertices, their connection, and color.

The remainder of the paper is organized as follows. In Section 2, detailed data description of 3D video is given. In Section 3, the modified shape distribution algorithm is described for stable shape feature extraction. Then, the algorithm for motion segmentation using the extracted feature vectors is explained in Section 4. Section 5 describes the algorithm for similar motion retrieval based on DP matching. Section 6 demonstrates the experimental results and concluding remarks are given in Section 7.

2. DATA DESCRIPTION

The 3D video data in the present work were obtained employing the system developed in [4]. They were generated from multiple view images taken with 22 synchronous cameras. The 3D object modeling is based on the combination of volume intersection and stereo matching [4].

Similar to 2D video, 3D video is composed of a consecutive sequence of "frames." Each frame of 3D video is represented as a polygon mesh model. Namely, each frame is expressed by three kinds of data as shown in Figure 1: coordinates of vertices, their connection (topology), and color.

The most significant feature in 3D video is that each frame is generated regardless of its neighboring frames. This is because of the nonrigid nature of human body and clothes. Therefore, the number of vertices and topology differ frame by frame, which makes it very difficult to search the correspondent vertices or patches among frames. Although Matsuyama et al. have been developing a deformation algorithm for dynamic 3D model generation [22], the number of vertices and topology needs to be refreshed every few frames.

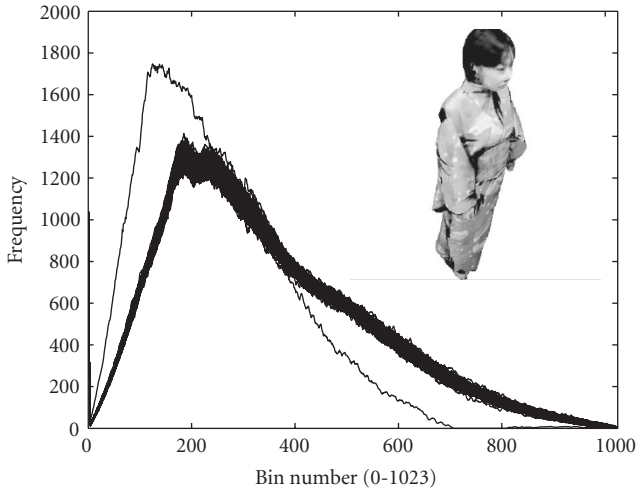


Figure 2: Thirty histograms for the same 3D model (shown on the upper side) using the original shape distribution [24]. Generated histograms have some deviation even for the same 3D model.

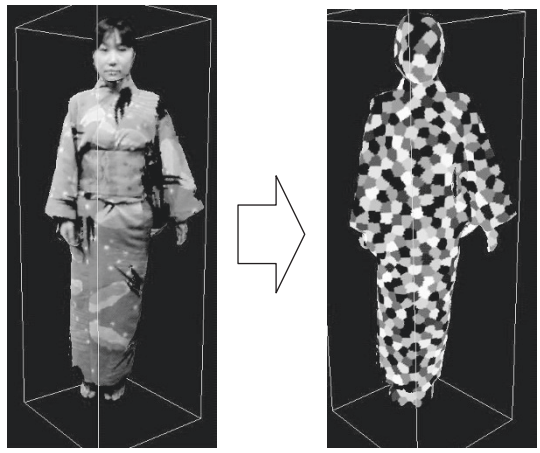


Figure 3: Concept of modified shape distribution. Vertices of 3D model are firstly clustered into 1024 groups by vector quantization in order to scatter representative vertices uniformly on 3D model surface.

3. SHAPE FEATURE EXTRACTION: MODIFIED SHAPE DISTRIBUTION

With regard to feature extraction from 3D models, a number of techniques have been developed aiming at static 3D model retrieval [23]. Among the feature extraction algorithms, shape distribution [24] is known as one of the most effective methods. In the original shape distribution algorithm [24], a number of points (e.g., 1024) were randomly sampled among the vertices of the 3D model surface and distance between all possible combinations of points was calculated. Then, a histogram of distance distribution was generated as a feature vector to express the shape characteristics of a 3D model. The shape distribution algorithm has a virtue of robustness to objects' rotation, translation, and so on.

However, histograms using the original shape distribution cannot be generated stably because of the random sampling of the 3D surface. Figure 2 shows 30 histograms generated for the same 3D model selected from our 3D video. The histograms were generated by randomly sampling 1024 vertices and setting the number of bins of the histogram as 1024 (dividing the range between maximum and minimum values in distance into 1024). It is observed that the shapes of the histograms fluctuate and sometimes a totally different histogram is obtained. In [24], deviation in the histograms was not so significant because rough shape feature extraction was pursued for similar shape retrieval of static 3D models. On the other hand, in our case, it is required to clarify a slight shape difference among frames in 3D video.

Therefore, we have modified the original shape distribution algorithm for more stability. Since vertices are mostly uniform on the surface in our 3D models, they are firstly clustered into 1024 groups based on their 3D spatial distribution employing vector quantization as shown in Figure 3. The centers of mass of the clusters are used as representative points for distance histogram generation. Although such clustering process is computationally expensive, it needs to be carried out only once in generating the histograms (feature vectors), and all the processings that follow are based on the extracted feature vectors. Therefore, the computational cost for clustering can be neglected. As a result, representative points are distributed uniformly and generation of stable histograms has been made possible. In our algorithm, the number of bins is set to 1024. After obtaining histograms, smoothing (moving average) is applied to them to remove noise by taking the average of the values in $-2\sim+2$ bins as shown in (1),

$$b_i' = \frac{b_{i-2} + b_{i-1} + b_i + b_{i+1} + b_{i+2}}{5}, \tag{1}$$

where b_i represents the ith element of the histogram and b_i' is that after the smoothing process. By using modified shape distribution, identical histograms can always be obtained for the same 3D model.

4. MOTION SEGMENTATION

In motion segmentation, for dance sequences in particular, motion speed is an important factor. When a person changes motion type or motion direction, the motion speed becomes small temporarily. More importantly, motion is paused for a moment to make the dance look lively. Such moments can be regarded as segmentation points.

Searching the points when the motion speed becomes small is achieved by looking for local minima in the degree of motion. From this point of view, our approach is similar to [8, 9]. The difference is that the degree of motion is calculated in the feature vector space since the movement of feature points of human body in 3D video is not clear as compared to motion capture data. Namely, the distance between the feature vectors of successive frames is utilized to express the degree of motion. In addition, one-dimensional

data of degree of motion goes thorough a further smoothing filter.

In [8], the extracted local minima in motion speed were verified whether they were truly segmentation boundaries or not by thresholding. This verification process is important to make the system robust to noise. The local minimum values should be lower than a predefined threshold value and the local maximum values between the local minima should be higher than another threshold. In this respect, threshold optimization depending on input data was still required in [8]. In our scheme, local minima are regarded as segmentation boundaries when the two local maxima on both sides of the local minimum value (D_{lmin}) are greater than $1.01 \times D_{\mathrm{lmin}}$. Since the verification is relative, it is robust to data variation and no empirical decision is required.

5. MATCHING BETWEEN MOTION CLIPS

In this paper, example-based queries are employed. A clip from a certain 3D video is given as a query and similar motion is searched from the other clips in the database. The performers in the query and the candidate clips do not necessarily have to be the same due to the robust shape feature representation by the modified shape distribution. However, since the shape distribution algorithm extracts the global shape feature, it is not eligible for searching motion clips with totally different types of clothes. For instance, a motion clip with casual cloth and that with Japanese *kimono* would be regarded as totally different motion sequences.

DP matching [20, 21] is utilized to calculate the similarity between the query and candidate clips. DP matching is a well-known matching method between time-inconsistent sequences, which has been successfully used in speech [25, 26], computer vision [27], and so forth.

A 3D video sequence in a database (Y) is assumed to be divided into segments properly in advance according to Section 4. Assume that the feature vector sequences of the query (Q) and the ith clip in Y, $Y^{(i)}$, are denoted as follows:

$$\begin{aligned} Q &= \{q_1, q_2, \ldots, q_s, \ldots, q_l\}, \\ Y^{(i)} &= \{y_1^{(i)}, y_2^{(i)}, \ldots, y_t^{(i)}, \ldots, y_m^{(i)}\}, \end{aligned} \tag{2}$$

where q_s and $y_t^{(i)}$ are the feature vectors of the sth and tth frames in Q and $Y^{(i)}$, respectively. Besides, l and m represent the number of frames in Q and $Y^{(i)}$.

Let us define $d(s,t)$ as the Euclidean distance between q_s and $y_t^{(i)}$ as in (3),

$$d(s,t) = ||q_s - y_t^{(i)}||. \tag{3}$$

Then, the dissimilarity (D) between the sequences Q and $Y^{(i)}$ is calculated as

$$D(Q, Y^{(i)}) = \frac{\mathrm{cost}(l,m)}{\sqrt{l^2 + m^2}}, \tag{4}$$

Table 1: Summary of 3D video utilized in experiments. Sequence #1 and sequences #2-1∼#2-3 are Japanese traditional dances called *bon-odori* and sequence #3 is a Japanese warmup dance. Sequences #2-1∼#2-3 are identical but are performed by different persons.

Sequence	# 1	# 2-1	# 2-2	# 2-3	# 3
Number of frames	173	613	612	616	1,981
Number of vertices (average)	83 k	17 k	17 k	17 k	17 k
Number of patches (average)	168 k	34 k	34 k	34 k	34 k
Resolution	5 mm	10 mm	10 mm	10 mm	10 mm
Frame rate	10 frames/s				

where the cost function $\mathrm{cost}(s,t)$ is defined as in the following equation:

$$\mathrm{cost}(s,t) = \begin{cases} d(1,1) & \text{for } l = m = 1, \\ d(s,t) + \min\{\mathrm{cost}(s,t-1), & \\ \qquad \mathrm{cost}(s-1,t), & \\ \qquad \mathrm{cost}(s-1,t-1)\} & \text{otherwise.} \end{cases} \tag{5}$$

Here, symbols of Q and $Y^{(i)}$ are omitted in $d(s,t)$ and $\mathrm{cost}(l,m)$ for simplicity. Since the cost is a function of the sequence lengths, $\mathrm{cost}(l,m)$ is normalized by $\sqrt{(l^2 + m^2)}$. The lower the D is, the more similar the sequences are.

6. EXPERIMENTAL RESULTS

In our experiments, five 3D video sequences generated by the system developed in [4] were utilized. The parameters of the data are summarized in Table 1. Sequences #1 and #2-1∼#2-3 are Japanese traditional dances called *Bon Odori* and sequence #3 is a Japanese warming-up dance. Sequences #2-1∼#2-3 are identical but performed by different persons. The frame rate was 10 frames/s. For the detailed content of 3D video, please see Figure 4 for sequence #1 and Figure 7 for #2-1. In sequences #2-1∼#2-3, the motion sequence in Figure 7 is repeated approximately three times.

6.1. Motion segmentation

In the experiment, the motion of "standing still" in the first tens of frames of each sequence was extracted manually in advance and neglected in the processing. Even when the dancer in 3D video is standing still, human body sways slightly, in which it is difficult to define segmentation boundaries.

Figure 4 demonstrates the subjective segmentation results for sequence #1 by eight volunteers. They were asked to define motion boundaries without any instruction or others' segmentation results. In this experiment, when four (50%) or more subjects voted for the same points, the segmentation boundaries were defined. The results were used for evaluation. For sequences #2-1∼#2-3 and #3, the segmentation boundaries were defined by the authors.

The segmentation results for the sequence #1 are illustrated in Figure 5. The ordinate represents the distance

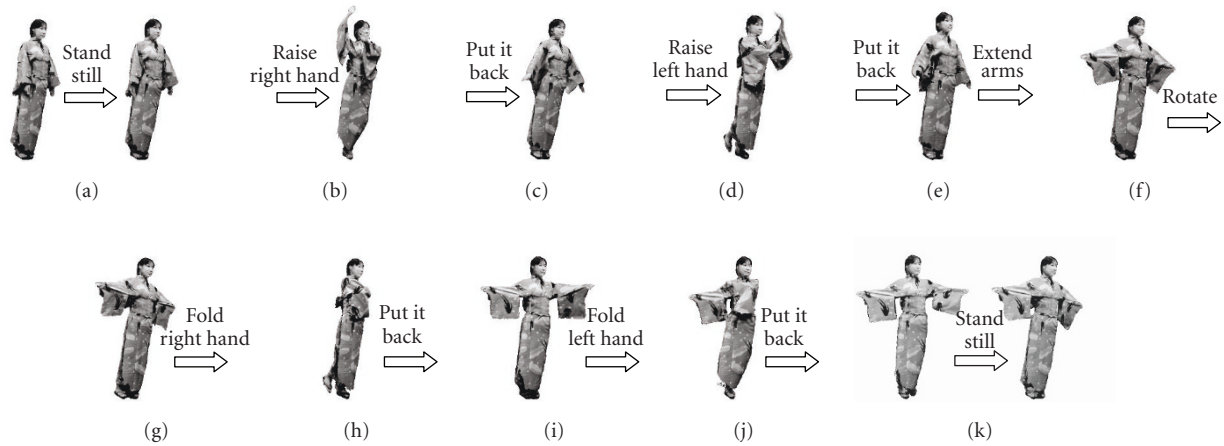


Figure 4: Subjective segmentation results for sequence #1 by eight volunteers.

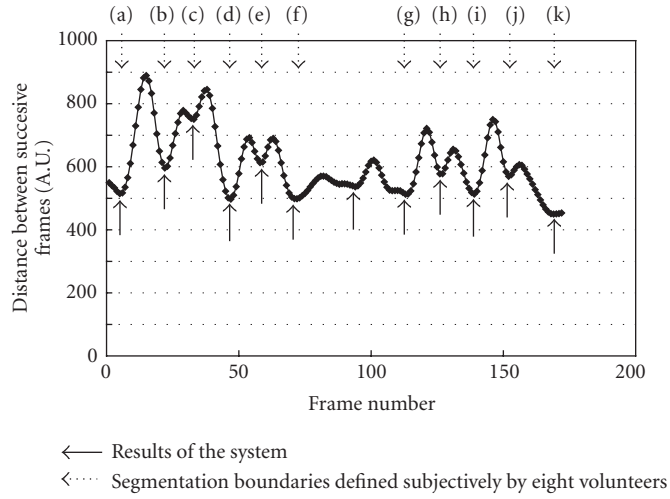


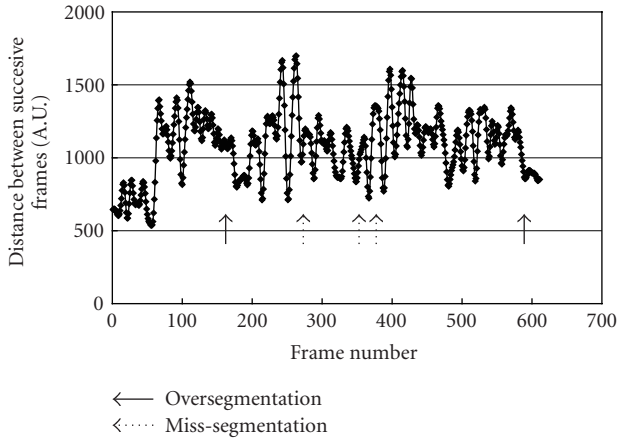
Figure 5: Comparison of subjectively defined segmentation points and results of our system for sequence #1. Dotted arrows from (a) to (k) represent the segmentation boundaries defined subjectively by eight volunteers. Solid arrows are the results of our system.

between histograms of successive frames. The dotted arrows from (a) to (k) represent the subjectively defined segmentation points shown in Figure 4. The solid arrows are the results of our system. There was only one over-segmentation. In addition, no miss-segmentation was detected. The over-segmentation between (f) and (g) was due to the fact that the pivoting foot was changed while the dancer was rotating and motion speed decreased temporarily (see Figure 9(a)).

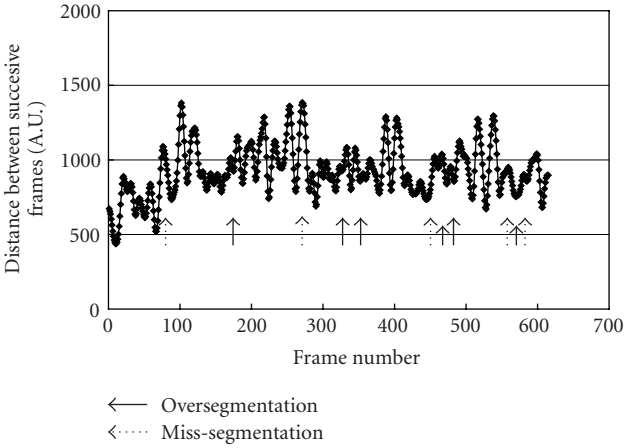
As other examples, segmentation results for sequences #2-1∼#2-3 are shown in Figure 6. The meanings of arrows are different from those in Figure 5. Solid arrows represent over-segmented points, and dotted arrows are miss-segmented points. The other local minima points coincided with authors' definition of segmentation boundaries. It is observed that the distances between the feature vectors of successive frames for sequences #2-1∼#2-3 are larger than those for sequence #1. This is because the dancer in sequence #1 wears *kimono* and motion in feet is not sensed very much.

The first 14 segmentation points (approximately, out of the 210 frames) obtained from sequence #2-1 are shown in Figure 7. It is observed that the 3D video sequence is divided into small but meaningful segments. There was only one over-segmentation, which is shown with the cross, and no miss-segmentation for the period. The precision and recall rates for sequence #2-1 were 95% and 93%, respectively (see Table 2 for more details).

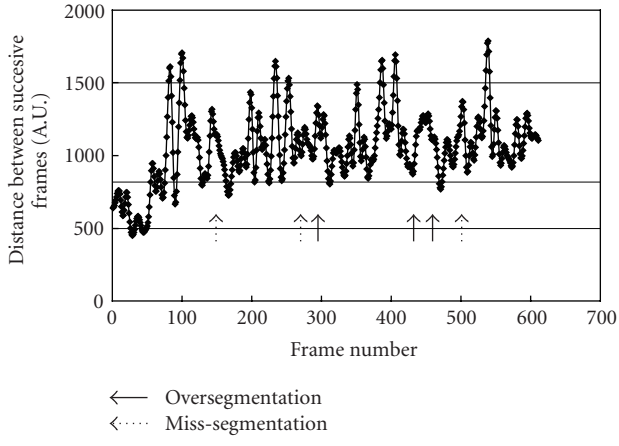
In our algorithm, only the distance between two successive frames is considered. Figure 8 shows the precision and recall rates when more neighboring frames are involved in the distance calculation using sequence #2-1. As the number of frames increases, recall rate is slightly improved while precision rate declines. This is because involving more neighboring frames in calculating the degree of motion corresponds



(a)



(b)



(c)

Figure 6: Segmentation results for sequences #2-1∼#2-3: (a) #2-1, (b) #2-2, (c) #2-3. The meanings of arrows are different from Figure 5. Solid arrows represent oversegmented points, and dotted arrows are miss-segmented points. The other local minima points coincided with authors' definition of segmentation boundaries.

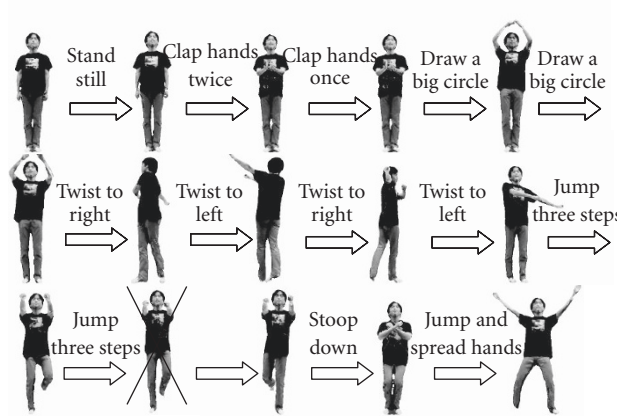


Figure 7: First 14 segmentation points of sequence #2-1. Image with cross-stands for oversegmentation.

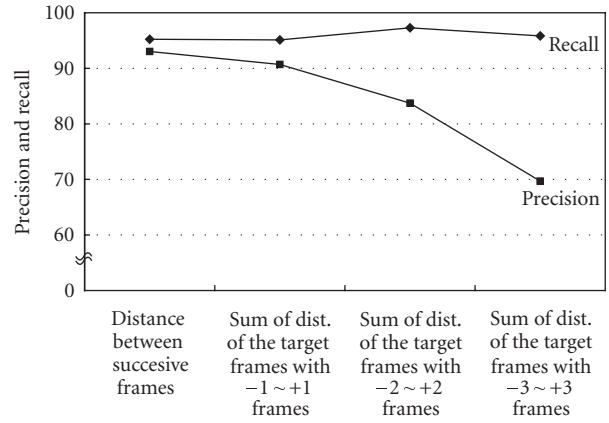


Figure 8: Precision and recall rates when the number of neighboring frames involved in calculation of degree of motion was changed. Sequence #2-1 was used.

Table 2: Performance summary of motion segmentation.

Sequence	# 1	# 2-1	# 2-2	# 2-3	# 3	Total
A: number of relevant records retrieved	11	40	42	34	124	251
B: number of irrelevant records retrieved	1	2	3	6	11	23
C: number of relevant records not retrieved	0	3	3	8	25	39
Precision: A/(A+B)	92	95	93	80	92	92
Recall: A/(A+C)	100	93	93	85	83	87

to neglecting small or quick motion. Our 3D video was captured at 10 frames/s. In such a low-frame rate case, calculating the distance between only the successive frames yields the best performance.

Table 2 summarizes the motion segmentation performance. The numbers of segmentation boundaries for #2-1∼ #2-3 are not the same because each dancer made some

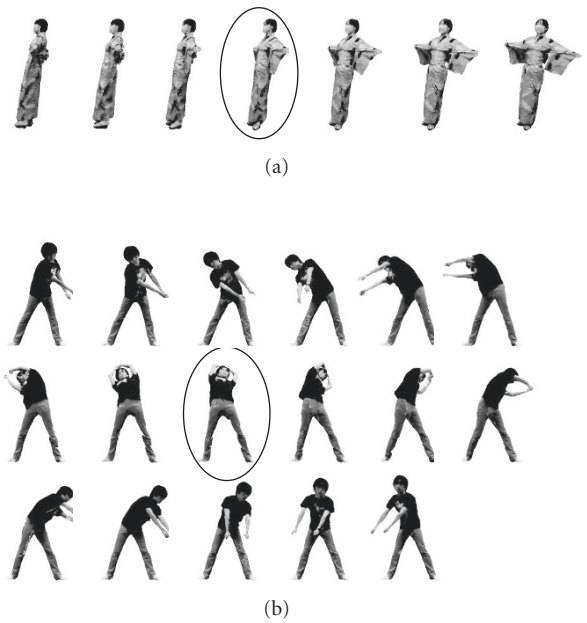


Figure 9: Examples of oversegmentation: (a) when changing pivoting foot; (b) when drawing a big circle by arms. Detected oversegmentation points are shown with circles.

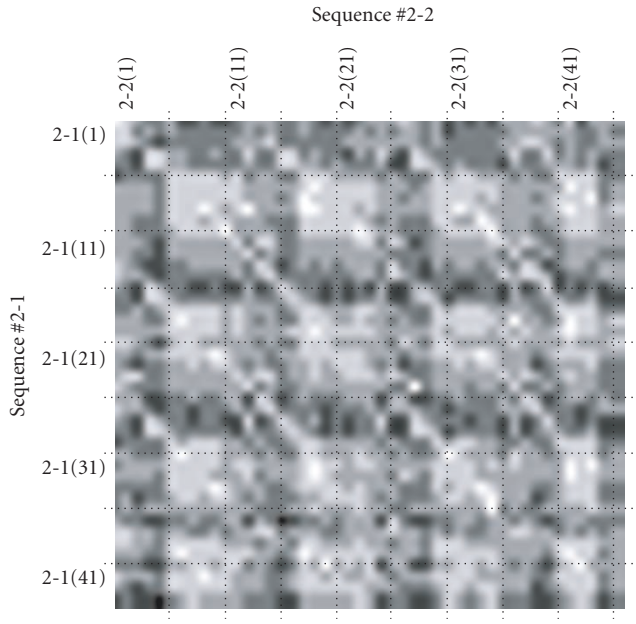


Figure 10: Matrix representing results of similarity evaluation between sequences #2-1 and #2-2. The whiter the color is, the more similar the two clips are.

mistakes. There are only a few miss- and over-segmentations per minute. Since sequence #3 contains more complicated motion than the others, which is hard to detect, the number of miss-segmentations is larger than the other sequences.

Most of the miss-segmentations were caused because the dancer did not pause properly even when the motion type changed. On the other hand, over-segmentation arose when the motion speed was decreased for motion transitions such as changing pivoting foot (Figure 9(a)) and changing the motion direction without changing the meaning of motion (Figure 9(b)). To resolve the problem, high-level motion observation may be needed.

6.2. Similar motion retrieval

In similar motion search, motion clips which are obtained by segmenting the sequences are handled as minimum units for computational efficiency. To demonstrate the retrieval performance itself, the miss- and over-segmentations in our motion segmentation results were corrected manually in advance. The motion definitions of the segmented clips after the correction in sequences #2-1 and #2-2 are shown in Table 3.

Figure 10 demonstrates the matrix representing the similarity evaluation score among clips in sequences #2-1 and #2-2. The brighter the color is, the more similar the two clips are. Although the dancers are different in sequences #2-1 and #2-2, it is observed that similar clips yield larger similarity score (smaller dissimilarity score D in (4)), showing the feasibility of our modified shape distribution-based retrieval.

Figure 11 shows an example of similar motion retrieval results. A motion clip of "drawing a big circle by hands (clip #2-$2^{(4)}$)" in sequence #2-2 was used as a query and similar motion was searched from clips in sequence #2-1. Figures 11(b)–11(g) demonstrate the top six most similar clips retrieved from sequence #2-1. It is demonstrated that similar motion is successfully retrieved even though the numbers of frames and posture of the 3D models are inconsistent with those in the query. In this case, all the relevant clips are retrieved. It has been confirmed that our retrieval system performs quite well for other queries.

Table 4 summarizes the retrieval performance using sequences #2-1 ~ #2-3. In the experiment, each clip from sequences shown in the column was used as a query. And the clips from the sequences shown in the row were used as candidates. The query itself was not included in candidates. The performance was evaluated by the method employed in [24]. The "first tier" in Table 4(a) demonstrates the averaged percentage of the correctly retrieved clips in the top k highest similarity score clips, where k is the number of the ground truth of similar motion clips defined by the authors. An ideal matching would give no false positives and would return a score of 100%. The "second tier" in Table 4(b) gives the same type of result, but for the top $2 \times k$ highest similarity score clips. The "nearest neighbor" in Table 4(c) shows the percentage of the test in which the retrieved clip with the highest score was correct. It is demonstrated that 56%~85% of similar motion clips are included in the first tier and more than 80% (82%~98%) of clips are correctly retrieved in the second tier. Besides, accuracy of nearest neighbor is 57%~98%. Therefore, it is observed that most of the similar motion can be found in the second tier. It is a rather good performance considering that only such low-level feature as the modified shape distribution is utilized in the matching.

Table 3: Motion definitions of clips after the correction: (a) sequence #2-1; (b) sequence #2-2.

(a)

Clip ID	Frames	Motion type
2-1$^{(1)}$	0–55	Stand still
2-1$^{(2)}$	56–71	Clap hands twice
2-1$^{(3)}$	72–82	Clap hands once
2-1$^{(4)}$	83–99	Draw a big circle
2-1$^{(5)}$	100–117	Draw a big circle
2-1$^{(6)}$	118–127	Twist to right
2-1$^{(7)}$	128–136	Twist to left
2-1$^{(8)}$	137–146	Twist to right
2-1$^{(9)}$	147–155	Twist to left
2-1$^{(10)}$	156–177	Jump three steps
2-1$^{(11)}$	178–192	Jump three steps
2-1$^{(12)}$	193–204	Stoop down
2-1$^{(13)}$	205–203	Jump and spread hands
2-1$^{(14)}$	204–224	Clap hands twice
2-1$^{(15)}$	225–234	Clap hands once
2-1$^{(16)}$	235–250	Draw a big circle
2-1$^{(17)}$	251–269	Draw a big circle
2-1$^{(18)}$	270–280	Twist to right
2-1$^{(19)}$	281–287	Twist to left
2-1$^{(20)}$	288–300	Twist to right
2-1$^{(21)}$	301–306	Twist to left
2-1$^{(22)}$	307–325	Jump three steps
2-1$^{(23)}$	326–347	Jump three steps
2-1$^{(24)}$	348–355	Stoop down
2-1$^{(25)}$	356–366	Jump and spread hands
2-1$^{(26)}$	367–374	Clap hands twice
2-1$^{(27)}$	375–387	Clap hands once
2-1$^{(28)}$	388–403	Draw a big circle
2-1$^{(29)}$	404–421	Draw a big circle
2-1$^{(30)}$	422–433	Twist to right
2-1$^{(31)}$	433–441	Twist to left
2-1$^{(32)}$	442–451	Twist to right
2-1$^{(33)}$	452–460	Twist to left
2-1$^{(34)}$	460–480	Jump three steps
2-1$^{(35)}$	481–501	Jump three steps
2-1$^{(36)}$	502–518	Jump and spread hands
2-1$^{(37)}$	519–527	Clap hands twice
2-1$^{(38)}$	528–537	Clap hands once
2-1$^{(39)}$	538–546	Twist to right
2-1$^{(40)}$	547–556	Twist to left
2-1$^{(41)}$	557–563	Twist to right
2-1$^{(42)}$	564–575	Twist to left
2-1$^{(43)}$	576–607	Stop in the middle of jumping
2-1$^{(44)}$	608–612	Stand still

(b)

Clip ID	Frames	Motion type
2-2$^{(1)}$	0–44	Stand still
2-2$^{(2)}$	45–61	Clap hands twice
2-2$^{(3)}$	62–71	Clap hands once
2-2$^{(4)}$	72–89	Draw a big circle
2-2$^{(5)}$	90–107	Draw a big circle
2-2$^{(6)}$	108–119	Twist to right
2-2$^{(7)}$	120–128	Twist to left
2-2$^{(8)}$	129–135	Twist to right
2-2$^{(9)}$	136–145	Twist to left
2-2$^{(10)}$	146–166	Jump three steps
2-2$^{(11)}$	167–181	Jump three steps
2-2$^{(12)}$	182–191	Stoop down
2-2$^{(13)}$	192–203	Jump and spread hands
2-2$^{(14)}$	204–214	Clap hands twice
2-2$^{(15)}$	215–224	Clap hands once
2-2$^{(16)}$	225–242	Draw a big circle
2-2$^{(17)}$	243–260	Draw a big circle
2-2$^{(18)}$	261–270	Twist to right
2-2$^{(19)}$	271–280	Twist to left
2-2$^{(20)}$	281–286	Twist to right
2-2$^{(21)}$	287–311	Twist to left
2-2$^{(22)}$	312–320	Undefined motion (mistake)
2-2$^{(23)}$	321–332	Jump three steps
2-2$^{(24)}$	333–344	Jump three steps
2-2$^{(25)}$	345–356	Stoop down
2-2$^{(26)}$	357–366	Jump and spread hands
2-2$^{(27)}$	367–372	Clap hands twice
2-2$^{(28)}$	373–392	Clap hands once
2-2$^{(29)}$	393–412	Draw a big circle
2-2$^{(30)}$	413–423	Draw a big circle
2-2$^{(31)}$	424–430	Twist to right
2-2$^{(32)}$	431–439	Twist to left
2-2$^{(33)}$	440–448	Twist to right
2-2$^{(34)}$	449–470	Twist to left
2-2$^{(35)}$	471–488	Jump three steps
2-2$^{(36)}$	489–499	Jump three steps
2-2$^{(37)}$	500–508	Stoop down
2-2$^{(38)}$	509–518	Jump and spread hands
2-2$^{(39)}$	519–529	Clap hands twice
2-2$^{(40)}$	530–544	Clap hands once
2-2$^{(41)}$	545–556	Draw a big circle
2-2$^{(42)}$	557–565	Twist to right
2-2$^{(43)}$	566–572	Twist to left
2-2$^{(44)}$	573–586	Twist to right
2-2$^{(45)}$	587–601	Twist to left
2-2$^{(46)}$	602–612	Stop in the middle of drawing a big circle

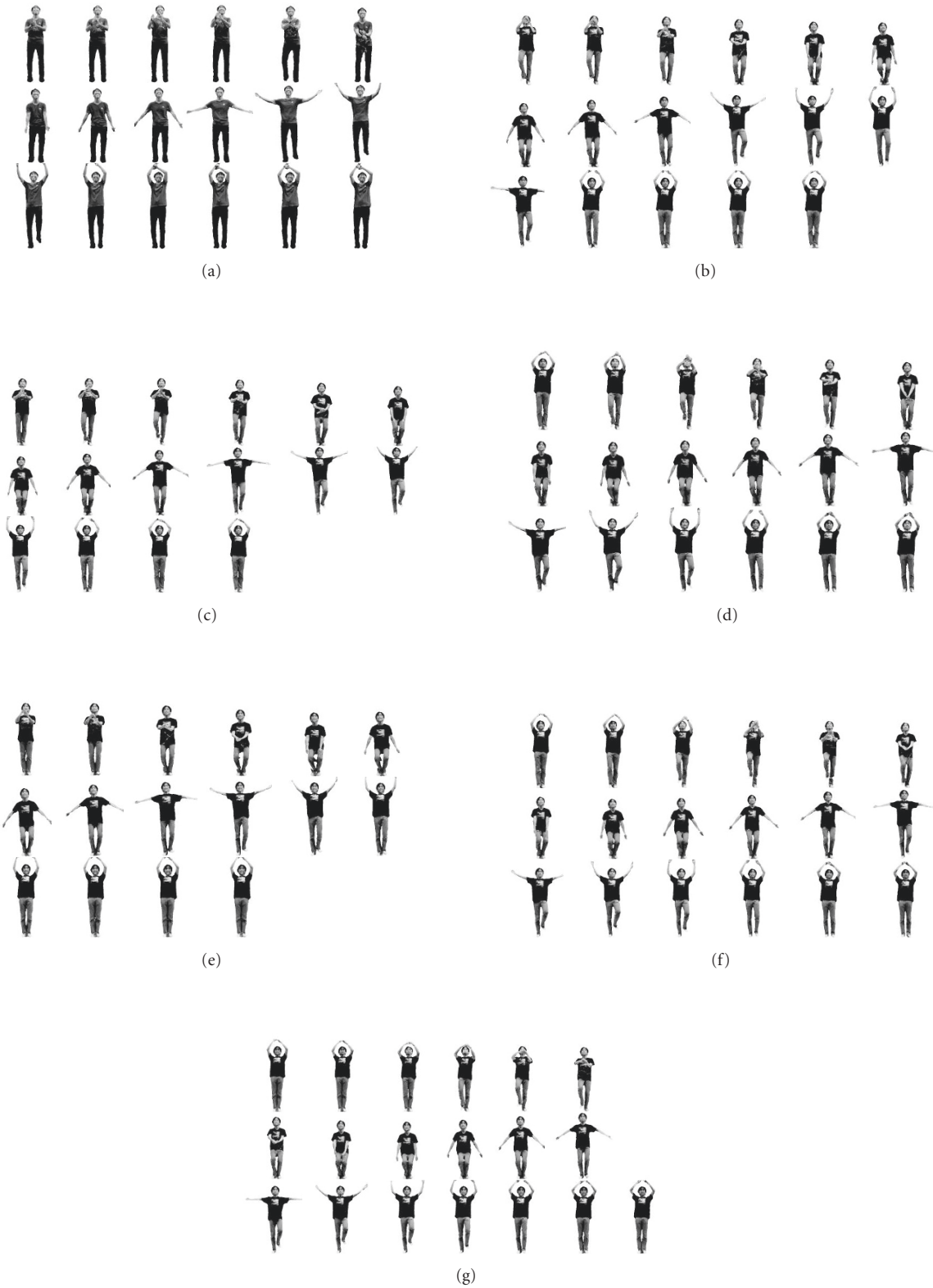


Figure 11: Experimental results for 3D video retrieval using motion of "drawing a big circle by hands": (a) query clip from sequence #2-2 (clip #2-2$^{(4)}$); (b) the most similar clip in sequence #2-1 (clip #2-1$^{(4)}$); (c) the second most similar clip (clip #2-1$^{(28)}$); (d) the third most similar clip (clip #2-1$^{(5)}$); (e) the fourth most similar clip (clip #2-1$^{(16)}$); (f) the fifth most similar clip (clip #2-1$^{(29)}$); (g) the sixth most similar clip (clip #2-1$^{(17)}$).

Table 4: Retrieval performance: (a) first tier, (b) second tier, (c) nearest neighbor. Query clip was generated from the sequence in the column and the clips from the sequences shown in the row were used as candidates. The query itself was not included in the candidate clips.

(a)

	# 2-1	# 2-2	# 2-3
# 2-1	80% (298/372)	63% (252/397)	70% (292/420)
# 2-2	62% (247/394)	63% (220/350)	63% (259/414)
# 2-3	57% (242/421)	56% (232/414)	85% (346/408)

(b)

	# 2-1	# 2-2	# 2-3
# 2-1	98% (366/372)	84% (335/397)	90% (378/420)
# 2-2	85% (335/394)	82% (287/350)	87% (360/414)
# 2-3	89% (374/421)	94% (390/414)	96% (392/408)

(c)

	# 2-1	# 2-2	# 2-3
# 2-1	98% (40/41)	76% (31/41)	90% (36/40)
# 2-2	57% (24/42)	62% (26/42)	62% (26/42)
# 2-3	67% (28/42)	69% (29/42)	90% (36/40)

Some false positives were detected due to the fact that the shape distribution is designed for extracting global shape features. Therefore, extracted sequential feature vectors tend to be affected by various factors such as difference in motion trajectories and physiques or clothes of the dancers. To enhance the retrieval performance, higher-level motion analysis is needed.

7. CONCLUSIONS

3D video, which is generated using multiple view images taken with a lot of cameras, is attracting a lot of attention as a new multimedia technology. In this paper, key technologies for 3D video retrieval such as feature extraction, motion segmentation, and similarity evaluation have been developed. The development of these technologies for 3D video is much more challenging than those for motion capture data because localization and tracking of feature points are very difficult in 3D video. The modified shape distribution algorithm has been employed for stable feature representation of 3D models. Segmentation has been conducted analyzing the degree of motion calculated in the feature vector space. The proposed segmentation algorithm does not require any predefined threshold values in verification process and relies only on relative comparison, thus realizing robustness to data variation. The similar motion retrieval has been realized by DP matching using the feature vectors. We have demonstrated effective segmentation with the precision and recall rates of 92% and 87% on average, respectively. In addition, reasonable retrieval results have been demonstrated by experiments.

ACKNOWLEDGMENT

This work is supported by the Ministry of Education, Culture, Sports, Science and Technology of Japan under the "Development of Fundamental Software Technologies for Digital Archives" Project.

REFERENCES

[1] T. Kanade, P. Rander, and P. J. Narayanan, "Virtualized reality: constructing virtual worlds from real scenes," *IEEE Multimedia*, vol. 4, no. 1, pp. 34–47, 1997.

[2] S. Wurmlin, E. Lamboray, O. G. Staadt, and M. H. Gross, "3D video recorder," in *Proceedings of the 10th Pacific Conference on Computer Graphics and Applications*, pp. 325–334, Beijing, China, October 2002.

[3] T. Matsuyama, X. Wu, T. Takai, and T. Wada, "Real-time dynamic 3-D object shape reconstruction and high-fidelity texture mapping for 3-D video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 3, pp. 357–369, 2004.

[4] K. Tomiyama, Y. Orihara, M. Katayama, and Y. Iwadate, "Algorithm for dynamic 3D object generation from multiviewpoint images," in *Three-Dimensional TV, Video, and Display III*, vol. 5599 of *Proceedings of SPIE*, pp. 153–161, Philadelphia, Pa, USA, October 2004.

[5] Y. Ito and H. Saito, "Free-viewpoint image synthesis from multiple-view images taken with uncalibrated moving cameras," in *Proceedings of IEEE International Conference on Image Processing (ICIP '05)*, vol. 3, pp. 29–32, Genova, Italy, September 2005.

[6] H. Habe, Y. Katsura, and T. Matsuyama, "Skin-off: representation and compression scheme for 3D video," in *Proceedings of Picture Coding Symposium (PCS '04)*, pp. 301–306, San Francisco, Calif, USA, December 2004.

[7] K. Müller, A. Smolic, M. Kautzner, P. Eisert, and T. Wiegand, "Predictive compression of dynamic 3D meshes," in *Proceedings of IEEE International Conference on Image Processing (ICIP '05)*, vol. 1, pp. 621–624, Genova, Italy, September 2005.

[8] T. Shiratori, A. Nakazawa, and K. Ikeuchi, "Rhythmic motion analysis using motion capture and musical information," in *Proceedings of IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI '03)*, pp. 89–92, Tokyo, Japan, July-August 2003.

[9] K. Kahol, P. Tripathi, and S. Panchanathan, "Automated gesture segmentation from dance sequences," in *Proceedings of the 6th IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 883–888, Seoul, Korea, May 2004.

[10] J. Barbič, A. Safonova, J.-Y. Pan, C. Faloutsos, J. K. Hodgins, and N. S. Pollard, "Segmenting motion capture data into distinct behaviors," in *Proceedings of Graphics Interface (GI '04)*, pp. 185–194, London, UK, May 2004.

[11] C. Lu and N. J. Ferrier, "Repetitive motion analysis: segmentation and event classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 2, pp. 258–263, 2004.

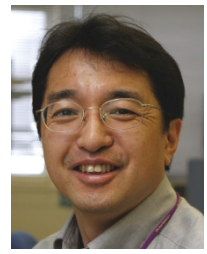
[12] W. Takano and Y. Nakamura, "Segmentation of human behavior patterns based on the probabilistic correlation," in *Proceedings of the 19th Annual Conference of the Japanese Society for Artificial Intelligence (JSAI '05)*, Kitakyushu, Japan, June 2005, 3F1-01.

[13] Y. Sakamoto, S. Kuriyama, and T. Kaneko, "Motion map: image-based retrieval and segmentation of motion data," in *Proceedings of ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pp. 259–266, Grenoble, France, August 2004.

[14] C.-Y. Chiu, S.-P. Chao, M.-Y. Wu, S.-N. Yang, and H.-C. Lin, "Content-based retrieval for human motion data," *Journal*

of Visual Communication and Image Representation, vol. 15, no. 3, pp. 446–466, 2004.
[15] M. Müller, T. Röder, and M. Clausen, "Efficient content-based retrieval of motion capture data," in *Proceedings of ACM SIGGRAPH*, pp. 677–685, Vienna, Austria, 2005.
[16] J. Xu, T. Yamasaki, and K. Aizawa, "3D video segmentation using point distance histograms," in *Proceedings of IEEE International Conference on Image Processing (ICIP '05)*, vol. 1, pp. 701–704, Genova, Italy, September 2005.
[17] J. Xu, T. Yamasaki, and K. Aizawa, "Effective 3D video segmentation based on feature vectors using spherical coordinate system," in *Meeting on Image Recognition and Understanding (MIRU '05)*, pp. 136–143, Hyogo, Japan, July 2005.
[18] T. Yamasaki and K. Aizawa, "Motion segmentation of 3D video using modified shape distribution," in *Proceedings of IEEE International Conference on Multimedia & Expo (ICME '06)*, Toronto, Ontario, Canada, July 2006.
[19] T. Yamasaki and K. Aizawa, "Similar motion retrieval of 3D video based on modified shape distribution," in *Proceedings of IEEE International Conference on Image Processing (ICIP '06)*, Atlanta, Ga, USA, October 2006.
[20] R. Bellman and S. Dreyfus, *Applied Dynamic Programming*, Princeton University Press, Princeton, NJ, USA, 1962.
[21] D. P. Bertsekas, *Dynamic Programming and Optimal Control (Volume One)*, Athena Scientific, Belmont, Mass, USA, 1995.
[22] T. Matsuyama, X. Wu, T. Takai, and S. Nobuhara, "Real-time 3D shape reconstruction, dynamic 3D mesh deformation, and high fidelity visualization for 3D video," *Computer Vision and Image Understanding*, vol. 96, no. 3, pp. 393–434, 2004.
[23] J. W. H. Tangelder and R. C. Veltkamp, "A survey of content based 3D shape retrieval methods," in *Proceedings of International Conference on Shape Modeling and Applications (SMI '04)*, pp. 145–156, Genova, Italy, June 2004.
[24] R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin, "Shape distributions," *ACM Transactions on Graphics*, vol. 21, no. 4, pp. 807–832, 2002.
[25] H. J. Ney and S. Ortmanns, "Dynamic programming search for continuous speech recognition," *IEEE Signal Processing Magazine*, vol. 16, no. 5, pp. 64–83, 1999.
[26] H. Ney and S. Ortmanns, "Progress in dynamic programming search for LVCSR," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1224–1240, 2000.
[27] A. A. Amini, T. E. Weymouth, and R. C. Jain, "Using dynamic programming for solving variational problems in vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 9, pp. 855–867, 1990.

Toshihiko Yamasaki was born in Japan in 1976. He received the B.S. degree in electronic engineering, the M.S. degree in information and communication engineering, and the Ph.D. degree from The University of Tokyo in 1999, 2001, and 2004, respectively. From April 2004 to October 2006, he was an Assistant Professor at Department of Frontier Informatics, Graduate School of Frontier Sciences, The University of Tokyo. Since October 2006, he has been an Assistant Professor at Department of Information and Communication Engineering, Graduate School of Information Science and Technology, The University of Tokyo. His current research interests include dynamic 3D mesh processing, wearable media, and analog VLSI design. He is a Member of IEEE, ACM, and so on.



Kiyoharu Aizawa received the B.E., the M.E., and the Dr.E. degrees in electrical engineering all from the University of Tokyo, in 1983, 1985, 1988, respectively. He is currently a Professor at the Department of Information and Communication Engineering of the University of Tokyo. He was a Visiting Assistant Professor at University of Illinois from 1990 to 1992. His current research interests are in image coding, image processing, image representation, video indexing, multimedia applications for wearable and ubiquitous environments, capturing and processing of person's experiences, multimedia processing for WWW, and computational image sensors. He received the 1987 Young Engineer Award and the 1990, 1998 Best Paper Awards, the 1991 Achievement Award, 1999 Electronics Society Award from the IEICE Japan, and the 1998 Fujio Frontier Award, the 2002 Best Paper Award from ITE Japan. He received IBM Japan Science Prize 2002. He serves as Associate Editor of IEEE Transactions on Circuit and Systems for Video Technology and is on the editorial board of the IEEE Signal Processing Magazine, the Journal of Visual Communications, and EURASIP Journal on Advances in Signal Processing. He has served for many national and international conferences including IEEE ICIP and he was the General Chair of SPIE VCIP99. He is a Member of IEEE, IEICE, ITE, ACM.