極座標表現を用いた形状特徴ベクトルによる3次元ビデオのセグメンテーション

徐 建鋒[†] 山崎俊彦[††] 相澤清晴[††]

複数のカメラを用いた多視点映像から生成される動的3次元オブジェクト（以後，本論文では「3次元ビデオ」と呼ぶ）は，実世界の3次元物体の詳細な情報を記録・再生できることから様々な研究が行われつつある．今後大規模な3次元ビデオのデータベースを構築し利活用していくためには，効率的な検索・編集技術の開発が求められる．本論文では検索・編集技術の開発に先立ち，まず被写体の時間的な動きの変化に着目した3次元ビデオのセグメンテーション技術を開発した．セグメンテーションとはビデオ・シーケンスを動きの意味のまとまりごとに分割する処理のことである．提案手法では極座標表現された3次元オブジェクトの頂点群から特徴ベクトルを生成することで，効率的でロバストなセグメンテーションを実現した．具体的には，ある基準点から各頂点への極座標 (r, θ, ϕ) を計算してそれぞれのパラメータのヒストグラムを特徴ベクトルとし，近傍フレーム間の特徴ベクトルどうしの距離がある条件を満たしたときにセグメンテーション位置であると判断した．これによって筆者らが以前に開発した3基準点によるヒストグラム法よりも，よりロバストなセグメンテーションが可能となった．3種類の3次元ビデオ・シーケンスに提案手法を適用した結果，適合率が0.77，再現率が0.95と良好な結果を得た．また，複数の被験者による主観評価の結果を統計的に処理し，セグメンテーションの性能を客観的に評価する指標もあわせて提案する．

Histogram-based Temporal Segmentation of 3D Video Using Spherical Coordinate System

Jianfeng Xu,[†] Toshihiko Yamasaki[††] and Kiyoharu Aizawa[††]

3D video, which is generated from multi-viewpoint images, has been drawing more attention because it can record and reproduce high-accuracy 3D information of the real-world objects. One of the most important issues in managing a large-scale database of 3D video archives is efficient retrieval for browsing, reusing, processing, and so on. Prior to retrieval systems, one has to solve the fundamental problem of temporal chopping of 3D videos into meaningful and manageable segments. In this paper, we have developed a robust and effective segmentation algorithm using histogram-based feature vector representation based on the spherical coordinate system. The segmentation algorithm developed in this paper has been applied to three different 3D video sequences and high *recall* and *precision* rates of 0.95 and 0.77, respectively, have been achieved in the best case. In addition, an objective performance evaluation method based on the statistical analysis of the subjective segmentation results is also proposed in this paper.

1. はじめに

近年，複数台のカメラで撮影した多視点画像から高精細な3次元ビデオを生成する研究がさかんに行われている[1)~4)]．3次元ビデオは従来のCGによる3次元オブジェクト合成やモーション・キャプチャによる3次元の動き情報取得に比べて，人間や動物など実世界の物体の姿・形・色などを忠実に記録・再現できるばかりでなく，その時間変化を追うことができる．また，3次元ビデオは撮影の際使用されたカメラ位置からのみでなく，任意視点からの視聴が可能である．そのため3次元ビデオは新しい映像表現として注目を浴びている．

3次元ビデオは新しい研究分野であるため，データの取得についてもまだ取り組むべき課題が多く，様々なシステムが研究されている．たとえば，Kanadeらは球状のスタジオに設置された複数台の同期カメラを使用して3次元ビデオ生成技術のプロトタイプを示し

† 東京大学大学院工学系研究科
Graduate School of Engineering, The University of Tokyo
†† 東京大学大学院新領域創成科学研究科
Graduate School of Frontier Sciences, The University of Tokyo

ている[1]．その後，Matsuyama らは視体積交差法によって 3 次元モデルの大まかな形状を取得したのち，動的弾性メッシュモデルという手法を導入することにより，滑らかなモデルの生成を可能にしている[3]．一方，Tomiyama らは大規模なスタジオを構築しており，視体積交差法とステレオ・マッチング法を組み合わせることで 2.5 mm〜5 mm の高精度な頂点解像度を実現している[4]．

参考文献 1)〜4) における 3 次元ビデオのデータは，一般的に 1 フレームずつ VRML によって記述され 3 角パッチによるメッシュモデルで表現される．すなわち，1 フレームのデータは 3 次元モデルの頂点位置・頂点どうしの結線情報・各頂点の色（または各 3 角パッチのテクスチャ）情報の 3 種類から構成されている．また，一般的に 3 次元ビデオのデータはフレームごとに独立に生成されるため，たとえ隣接するフレームどうしであっても頂点数や結線情報は異なるのが特徴である☆．

今後，大規模な 3 次元ビデオデータベースを構築して実用的に利活用できるようにするためには，取得ばかりでなく検索・編集技術の開発が必要不可欠である．効率良く検索や編集を行うためには，まず被写体の動きによって映像を細かく分割するセグメンテーション（動きの「分節化」とも呼ばれる）が重要な役割を果たす．たとえば，中澤らはモーションキャプチャ・データで取得された舞踊に対してセグメンテーションを行うことにより舞踊全体の構成を把握したり[6]，ロボットによる舞踊再現に利用したりしている[7]．また，様々な舞踊をセグメンテーションによって細分化し，舞踊譜に変換してデジタルアーカイブ化する試みも行われるなど[8]，セグメンテーションは様々な技術の前処理として重要な役割を果たしている．ここで特記すべきなのは，本論文で扱おうとしているセグメンテーションは被写体の動きの分節化のことであり，2 次元映像でさかんに研究されているシーン変化検出[9),10)] とは異なるということである．

これまで開発されてきた動きセグメンテーション技術は，主に 2 次元映像[11),12)] やモーションキャプチャ・データに関するもの[13)〜17)] である．2 次元映像のセグメンテーションにおいては，まず背景と動いている物体を分離する．その後，参考文献 11) では動いている物体のオプティカル・フローに対して特異値分解を施し，オプティカル・フローを主成分のみで表現する．動きの種類が変化するとその時点でオプティカル・フローの主成分が大きく変化するので，それを利用してセグメンテーションを行っている．また，参考文献 12) では動きの大きさが極少，かつ動きの方向変化が極大となる時点を探索することでセグメンテーションを実現している．

モーションキャプチャ・データに対するセグメンテーション技術もこれまで数多く提案されている[13)〜17)]．それは，関節やその他の特徴点の位置特定・動き追跡が容易に行えるからである．たとえば，参考文献 13) においては動きの大きさが極小になる時点を用いてセグメンテーションを行っている．運動力学的特徴量が極小となる時点を探索するという手法は参考文献 14) にも取り入れられている．また，動き予測誤差に基づく手法も特異値分解を用いたもの[15)]，最小二乗フィッティングを用いたもの[16)] が提案されている．さらには，隠れマルコフモデル[17)] や Gaussian Mixture Model [14)] を用いたモデルベースのアプローチも提案されている．

以上に紹介してきた 2 次元映像やモーションキャプチャ・データに比べて，3 次元ビデオのセグメンテーションはほとんど報告例がない．3 次元ビデオのセグメンテーションを行うにあたっての難しさはモーションキャプチャ・データのようにフレーム間の明確な対応が存在せず，関節やその他特徴的な点の位置特定および追跡が非常に困難であるということにある．これは前述のように 3 次元ビデオは原理的に 1 フレームずつ独立に生成されるため，たとえ隣り合うフレームどうしであっても頂点の数や接続関係がフレームごとに変化することによる．筆者らは 3 次元ビデオのセグメンテーションについて取り組み，これまでに基準点と 3 次元モデル頂点の距離ヒストグラムを用いたセグメンテーションを提案してきた[18),19)]．筆者らの知る限り 3 次元ビデオ・セグメンテーションとしてはこれらが初めての試みである．ヒストグラムを用いた手法は処理が簡単で大量のデータ処理にも適しており，またノイズの影響を受けにくいという特徴を持つ．また，3 つの頂点からヒストグラムを生成することで 3 次元モデルの頂点が基準となる原点を中心とした球上を動いた場合，距離に基づくヒストグラムにその影響が反映されないという問題を解決してきた[18)]．しかしその反面，3 つの基準点を自動的に決定するのが困難である，またオブジェクトの回転運動をうまく表現できない場合があるなどの問題点もあった．

本論文の目的は，よりロバストな特徴ベクトル抽出手法を開発しセグメンテーションの精度を向上させることである．そこで今回，3 次元オブジェクトの頂点

☆ Matsuyama らは頂点数や結線情報を保ったまま 3 次元ビデオを生成する技術の開発も行っているが[5)]，数フレームに 1 回の頻度で頂点数・結線情報は初期化する必要がある．

群を極座標表現に変換し，そこから特徴ベクトルを生成することで効率的でロバストなセグメンテーションを行う技術を開発した[19]．具体的には，あらかじめ定めておいた基準点から各頂点への極座標 (r, θ, ϕ) を計算して，それぞれのパラメータのヒストグラムを特徴ベクトルとした．そのうえで参考文献 18) と同様に近傍フレーム間の特徴ベクトルの距離がある条件を満たしたときにセグメンテーション位置であると判断した．これによって筆者らが以前に開発した 3 基準点によるヒストグラム法（Point Distance Histogram: PDH）よりも，ロバストなセグメンテーションが可能となった．

また，セグメンテーションのように人間が主観評価を行った場合でも被験者によって定義にばらつきが生じるものについて，いかに客観的に性能評価を行うかということも大きな問題である．従来の研究では，セグメンテーションの例のみを提示して客観的な性能評価を行っていないものや，人間による主観評価結果と比較しているものの被験者間の定義のばらつきをどう扱うかについては議論されていないものがほとんどであった．そこで，本論文では複数の被験者による主観評価の結果を統計的に処理し，セグメンテーションの性能を客観的に評価する手法もあわせて提案する[20]．3 種類の 3 次元ビデオ・シーケンスに対して本論文で開発したセグメンテーション手法を施し，提案する評価尺度で性能評価したところ，適合率が 0.77，再現率が 0.95 と良好な結果を得た．

本論文の構成は以下のようになっている．まず 2 章で筆者らが扱おうとしている 3 次元ビデオのデータ構造について述べる．3 章では 3 次元ビデオ・セグメンテーションのアルゴリズムについて詳しく説明する．4 章で評価手法について議論した後，5 章で実験結果を示す．6 章は結論と今後の課題である．

2. 3 次元ビデオのデータ構造

本論文で扱う 3 次元ビデオ・データは Tomiyama ら[4] によって多視点映像の処理により取得・生成されたもので，1 フレームずつ Virtual Reality Modeling Language（VRML）によってポリゴンメッシュモデルとして記述されている．**図 1** に 3 次元ビデオの例を示す．VRML は ISO/IEC 14772-1 に定義されている国際標準規格で，現在インターネット上などで手に入れられる 3 次元モデルの多くも VRML またはそれに類する言語によって記述されている．また，VRML は MPEG-4 の Animation Framework eXtension（AFX）でもサポートされている．

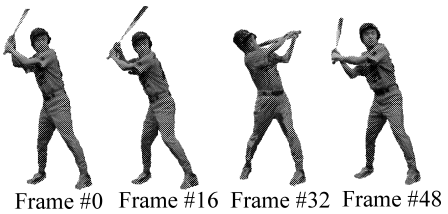

図 1 3 次元ビデオの例．ここでは 1 視点からの画像のみを示す
Fig. 1 Samples of 3D video. Images from a single viewpoint are shown.

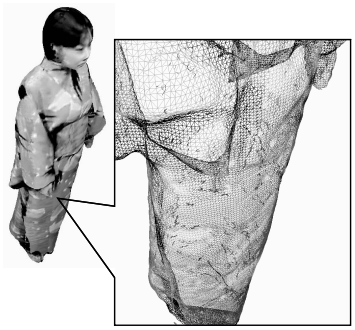

図 2 3 次元モデルの拡大図．3 次元モデルは頂点座標，それぞれの接続関係，各頂点の色の 3 つの情報から構成されている
Fig. 2 Detailed 3D model. 3D model consists of coordinates of vertices, their connection and color.

表 1 用いた 3 次元ビデオのデータ諸元
Table 1 Parameters of our test sequences.

Sequence Name	ダンス	バッティング	ピッチング
フレーム数	173	51	51
平均頂点数	83,477	64,254	65,516
平均 3 角パッチ数	168,460	128,524	131,052

1 フレーム内の 3 次元モデルはユークリッド座標系で定義された頂点群とそれぞれの結線関係，および各頂点の色の 3 種類のデータから構成されている（**図 2** 参照）．本論文で用いた 3 次元ビデオはダンス，ピッチング，バッティングの 3 種類のシーケンスで，それぞれ毎秒 10 フレームで生成されている．**表 1** にデータの諸元を示す．シーケンスの長さはそれぞれ 173，51，51 フレームである．また，3 次元ビデオは多視点映像から各フレームが独立に生成されるため，頂点数や結線情報などは隣り合うフレームどうしであっても異なるのが特徴である．

3. 3 次元ビデオのセグメンテーション

3.1 システムの概要

本論文では特に舞踊などの動きをターゲットにし，3 次元ビデオ中のオブジェクトの動きに基づいてセグメンテーションを行う．一般に，動きの種類や運動の方

向が変化するとき，動きは一時的に小さくなる．そこで，本論文では動きが大きな区間と小さな区間に分割することでセグメンテーションを実現する．この手法は，動きの大きさの変化に注目するという点で参考文献 12)〜14) などと同様の考え方である．しかし，ここで重要なのは 3 次元ビデオでは先に述べたとおり関節や特徴点の位置特定および追跡が困難な点である．そのため，提案方式ではまず各フレームのオブジェクトの形状・姿勢から特徴ベクトルを抽出し，特徴ベクトル空間で動きの大きさを解析する．すなわち，実空間でのオブジェクトの動きを理解・解析しなくてもセグメンテーションを行えることを目指した．この点が従来の動きの大きさを用いたセグメンテーション[12)〜14)]との大きな違いである．本論文では，ヒストグラム・ベースの特徴ベクトル抽出方式を開発した．ヒストグラムに基づく特徴量抽出はノイズに対してロバストで計算コストが低いという利点がある．提案手法による実験結果は 4 章で述べるように 8 名の主観評価に基づいたセグメンテーション結果を基に評価した．

3.2 特徴ベクトル生成

開発したアルゴリズムの一番大きな特徴は 3 次元オブジェクトの頂点座標を極座標を用いて表現することである．本手法を今後 Spherical Coordinate Histogram（SCH）法と呼ぶ．極座標表現で特に θ，ϕ のレンジを広くとるためには，座標系の原点は 3 次元オブジェクトの内部にあることが望ましい．そこで，式 (1) に示すようにすべてのフレームの頂点をまず初期フレームの重心位置を用いて平行移動する．

$$\begin{cases} x'_i(t) = x_i(t) - x_0 \\ y'_i(t) = y_i(t) - y_0 \\ z'_i(t) = z_i(t) - z_0 \end{cases} \tag{1}$$

ただし，

$$\begin{cases} x_0 = \frac{1}{N(0)} \sum_{i=0}^{N(0)-1} x_i(0) \\ y_0 = \frac{1}{N(0)} \sum_{i=0}^{N(0)-1} y_i(0) \\ z_0 = \frac{1}{N(0)} \sum_{i=0}^{N(0)-1} z_i(0) \end{cases} \tag{2}$$

とする．ここで，t は時間的なフレームのインデックスを，i は t 番目のフレーム内の頂点インデックスを，そして $N(t)$ は t フレーム目のモデルに含まれる頂点の数を表している．また x，y，z，x'，y'，z' はそれぞれ重心補正前後のユークリッド座標系での頂点座標を表す．なお，ここで注意すべきなのは頂点の平行移動はフレームごとの重心位置ではなく初期フレームの重心位置を用いて行っていることである．これによって隣接フレーム間の微妙な動きの変化を重心補正によって打ち消されてしまうのを防ぐ．

重心補正を行った後，3 次元オブジェクトの各頂点は以下の式により極座標 (r, θ, ϕ) に変換される（**図 3** 参照）．

$$\begin{cases} r_i(t) = \sqrt{x'^2_i(t) + y'^2_i(t) + z'^2_i(t)} \\ \theta_i(t) = sign(y'_i(t)) \cdot arccos\left(\frac{x'_i(t)}{\sqrt{x'^2_i(t) + y'^2_i(t)}}\right) \\ \phi_i(t) = arccos(\frac{z'_i(t)}{r_i(t)}) \end{cases} \tag{3}$$

ここで，$sign$ は sign 関数で，以下の式によって定義される．

$$sign(x) = \begin{cases} +1 & for\ x > 0 \\ 0 & for\ x = 0 \\ -1 & for\ x < 0 \end{cases} \tag{4}$$

この式において，r，θ，ϕ の範囲はそれぞれ $r \in [0, \infty)$，$\theta \in [-\pi, \pi)$，$\phi \in [0, \pi]$ である．

座標変換後，r，θ，ϕ それぞれのヒストグラムを生成し，それらを特徴ベクトルとして用いる．提案手法では bin 幅を一定とし，r の最大値までの範囲でヒストグラムを生成する．ここで，θ，ϕ に関してはとりうる値の範囲は有限なので bin 幅が決まれば自動的に bin 数も決定される．それに対し，r のとりうる範囲は理論上 $[0, \infty)$ なので bin 幅および bin 数をどのように決定するかが問題となる．一般的には r の最小値・最大値を用いて正規化するといった処理が行われることが多い．しかし，r の最大値・最小値が変化したということも動きのセグメンテーションにとっては重要な情報である場合も多い．また逆に正規化によってノイズの影響を強く受けてしまう場合もあるため（5 章に詳述），正規化処理は本手法には不適切である．あらかじめ非常に大きな値 R を考えてその中を一定の bin 数に分割することも考えられるが，r の最大値が

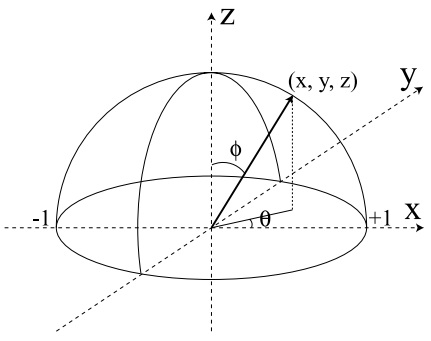

図 3 極座標変換
Fig. 3 Conversion to spherical coordinate system.

R に比べて非常に小さい場合，どのような動きをしても r のヒストグラムにその情報が反映されない状況に陥る．そこで，本手法ではあえて bin 幅だけを定めて bin 数はモデルの形状に合わせて可変とする．この場合 bin 数は理論上無限大になるが，実際は有限であるので処理は可能である．また，これによって r に関するヒストグラムの bin 数がフレームにより異なるが，フレーム間のヒストグラムの距離を計算する際には次節に述べる方法で対応する．

3 次元オブジェクトの動きはヒストグラムの変化として現れる．仮に 3 次元オブジェクトの動きが偶然に図 3 の原点を中心に回転するようなものであったとした場合，r のヒストグラムにはその影響が反映されないが，θ・ϕ のいずれかのヒストグラムには必ず反映される．なお，本提案手法は運動の角度情報を持っているので，文献 18) の PDH 法に比べて伝統芸能によく見られるような同一地点での回転などの動きも検出することができる（5 章に詳述）．

3.3 セグメンテーション位置の検出

セグメンテーション位置の検出は，隣接するフレーム間のヒストグラムの類似度を評価することで行う．まず，r，θ，ϕ それぞれについて以下のようにユークリッド距離を計算する．

$$
\begin{aligned}
d(r,t) &= \sqrt{\sum_{j=0}^{max(J_r(t),J_r(t+1))-1} \left(h^*_{j,r}(t+1) - h^*_{j,r}(t)\right)^2} \\
d(\theta,t) &= \sqrt{\sum_{j=0}^{J_\theta-1} \left(h_{j,\theta}(t+1) - h_{j,\theta}(t)\right)^2} \qquad (5) \\
d(\phi,t) &= \sqrt{\sum_{j=0}^{J_\phi-1} \left(h_{j,\phi}(t+1) - h_{j,\phi}(t)\right)^2}
\end{aligned}
$$

ただし，J_r，J_θ，J_ϕ はそれぞれの bin 数を表しており，$h_{j,r}(t)$，$h_{j,\theta}(t)$，$h_{j,\phi}(t)$ はそれぞれ r，θ，ϕ の t フレーム目のヒストグラムの j 番目の bin の要素の値を表している．また，$h^*_{j,r}$ は式 (6) のように定義される．

$$
h^*_{j,r}(t) = \begin{cases} h_{j,r}(t) & for\ j < J_r(t) \\ 0 & otherwise \end{cases} \qquad (6)
$$

これは r に関するヒストグラムのみ bin 数がフレームごとに異なるための措置である．3 つのパラメータのユークリッド距離を計算した後，以下の式によって隣接フレーム間の最終的な距離の定義とする．

$$
d(t) = \sqrt{d^2(r,t) + d^2(\theta,t) + d^2(\phi,t)} \qquad (7)
$$

表 2 実験に用いた bin 幅，閾値などのパラメータ
Table 2 Parameters in the experiments.

Parameter	SCH1	SCH2	PDH1	PDH2
$l(r)$	0.8	1.6	-	-
$l(\theta)$	0.05	0.1	-	-
$l(\phi)$	0.05	0.1	-	-
$l(DP)$	-	-	1.1	2.2
Th1	0.1	0.1	0.03	0.03
Th2	1.2	1.2	1.65	1.65
Th3	0.6	0.6	0.8	0.8
α	0.5	0.5	0.5	0.5

ここで，3 つのパラメータ r，θ，ϕ のセグメンテーションに対する影響力はそれぞれ異なる．そこで本論文では r，θ，ϕ それぞれに対する bin 幅を実験的に最適化し（**表 2** 参照），どれか 1 つのパラメータの変化に過敏に反応することのないような距離演算を行う．

セグメンテーション位置の検出は文献 18) で開発したものを用いる．この手法はただ単に隣接フレーム間の距離だけを見るのではなく，2 フレーム以上続けてある条件を満たした場合のみセグメンテーション位置であると判断するため，ノイズへの耐性が強くロバストな検出を行うことができる．式 (8)，(9) にその判定アルゴリズムを示す．

Abrupt transition:

$$
|d(t+1) - d(t)| > Th1 \qquad (8)
$$

Gradual transition:

$$
\begin{cases} \frac{d(t)}{\overline{d(k)}} > Th2 \quad and\ \frac{d(t+1)}{\overline{d(k)}} > Th2 \\ or \\ \frac{d(t)}{\overline{d(k)}} < Th3 \quad and\ \frac{d(t+1)}{\overline{d(k)}} < Th3 \end{cases} \qquad (9)
$$

ここで $\overline{d(k)}$ は直前のセグメンテーション位置から現フレーム位置まで（フレーム数を k とする）の隣接フレーム間距離の平均値を示す．また，$Th1$〜$Th3$ は閾値である．式 (8) は通常の映像でいうところのカット点など，急激にシーンが変化する点を検出する処理である．一方，式 (9) は同一シーン内での動きの意味の変化といった，なだらかな変化を検出するために用いる．$Th2$ は動きの変化量（すなわちフレーム間距離）が小さい区間が続いた後大きくなり始める時点を検出し，$Th3$ は逆に動きの変化量が大きい区間が続いた後小さくなり始める時点を検出する．

4. 評価手法

セグメンテーションの正解位置については，客観的に絶対正しい位置というものは存在しない．そのため，従来の研究では，セグメンテーションの例のみを提示して客観的な性能評価を行っていないものや，人間に

よる主観評価結果と比較しているものの被験者間の定義のばらつきをどう扱うかについては議論されていないものがほとんどであった．本論文では複数の被験者による主観評価の結果を統計的に処理し，セグメンテーションの性能を客観的に評価する手法を提案する．

まず，3 次元ビデオのセグメンテーションについて予備知識のない 8 人の被験者に個別に 3 次元ビデオを提示し，自由に視点を操作しながら主観的にセグメンテーションを行ってもらった．3 次元ビデオの閲覧回数に制限は設けなかった．そのため，視点に依存した「隠れ」の影響はなく，あらゆる視点を考慮した主観評価が行えたものとする．なお，主観評価によるセグメンテーションを行う際，被験者は予備知識および例の提示はまったく受けておらず，他の被験者のセグメンテーション結果もお互いに知らないまま実験を行った．

図 4 に 8 人の被験者によるダンス・シーケンスのセグメンテーション結果を示す．この時点で主観的セグメンテーションの結果には個数や位置にばらつきが存在することが分かる．さらに，3 次元ビデオの動きの変化の種類・大きさなどによって被験者が定義するセグメンテーション位置のばらつき方にも違いがある．そのため，本論文では主観的セグメンテーション結果の個数および位置のばらつきを統計的に処理し，主観評価のばらつき方を考慮に入れた性能評価手法を開発した．

まず最初に，図 4 の結果を筆者らの主観により同じ区切れに対して投票していると思われるものどうしにグループ分けする．この作業は各被験者がどの動き区切れ位置に投票したか，ということを調査するだけなのですべて正しくグループ分けされるものと仮定する（誰がこの作業を行っても同じグループ分けの結果が得られると仮定する）．その結果，**図 5** に示すようなセグメンテーション位置候補と各候補への投票数が明らかとなる．

その後，どれだけの投票数があれば（おおよその）正解位置として認めるかの閾値を定める．本論文では投票数 4 人以上，すなわち 50%以上の被験者による投票があった位置を正解位置であると定義した．ただし，この段階ではまだ正解のおおよその位置が定義されるだけで，正確な位置の定義はなされていないことに注意されたい．先に述べたように，主観評価の投票結果には位置のばらつきが存在する．さらに，動きの区切り位置が明確な場合には同じフレームに投票が集中し，また区切り位置が不明確な場合には投票位置は大きくばらつく．そこで，本論文では投票位置のばらつきにガウス分布を想定し，ばらつきの標準偏差を

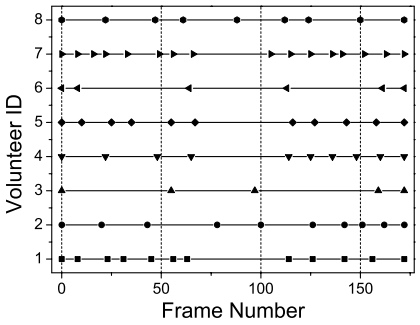

図 4 8 人の被験者によるダンス・シーケンスの主観的セグメンテーション結果．点の位置がそれぞれの被験者のセグメンテーション位置を示す

Fig. 4 Segmentation results by eight subjects for dance sequence.

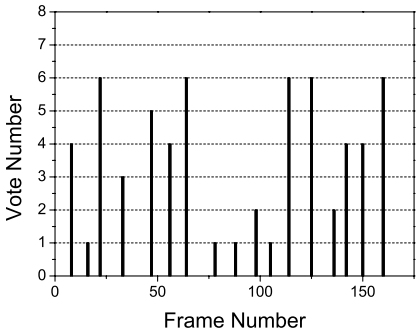

図 5 8 人の被験者によるそれぞれの動き区切り位置への投票数．動き区切り位置は著者らが図 4 より主観的に判断

Fig. 5 Vote number for each segmentation candidates based on subjective segmentation by 8 volunteers.

σ としたとき，筆者らが開発したシステムのセグメンテーション位置が平均位置から 3σ（99.7%）の範囲に収まっていれば正解であると判定する．以上の定義をもとに，以下示す適合率（Precision rate: P），再現率（Recall rate: R），F 値をもってセグメンテーションの精度評価を行う．

$$P = \frac{\text{正解の数}}{\text{検出された数}} \tag{10}$$

$$R = \frac{\text{正解の数}}{\text{正解の総数}} \tag{11}$$

$$F = \frac{1}{\alpha \cdot \frac{1}{P} + (1-\alpha) \cdot \frac{1}{R}},\ \alpha \in [0,1] \tag{12}$$

F 値は適合率と再現率の調和平均であり，適合率・再現率の両方が高いときのみ F 値も高くなることから，検出精度のトータルバランスを評価するのに用いられる．α は適合率と再現率の重みで，一般的には 0.5 に設定される．

5. 実験結果

実験では，2 章で説明した 3 つのシーケンス（ダンス，バッティング，ピッチング）を順につなげ，トータル 275 フレームとした．すなわち，急峻なシーケンスの切り替わり位置は 2 つ存在する．実験に用いたパラメータを表 2 に示す．なお，r，θ，ϕ の bin 幅は SCH1 においてはそれぞれ 0.8，0.05，0.05 とした．また，SCH2 の bin 幅はそれぞれ 0.16，0.1，0.1 と SCH1 の 2 倍に設定した．

まず，bin 幅を固定にした場合と可変にした場合のヒストグラムの違いを**図 6** に示す．3 次元オブジェクトの形状が互いに似ていても bin 幅を可変（bin 数を固定し，r の最大値・最小値を用いて正規化）にした場合はバット位置の微妙な変化の影響を受けてヒストグラムの形状が大きく異なっているのに対し，bin 幅を固定した場合は（bin 幅を 0.8 に設定）安定してヒストグラムが得られていることが分かる．このことから，bin 幅を固定した方がセグメンテーションには適していることが明らかとなった．

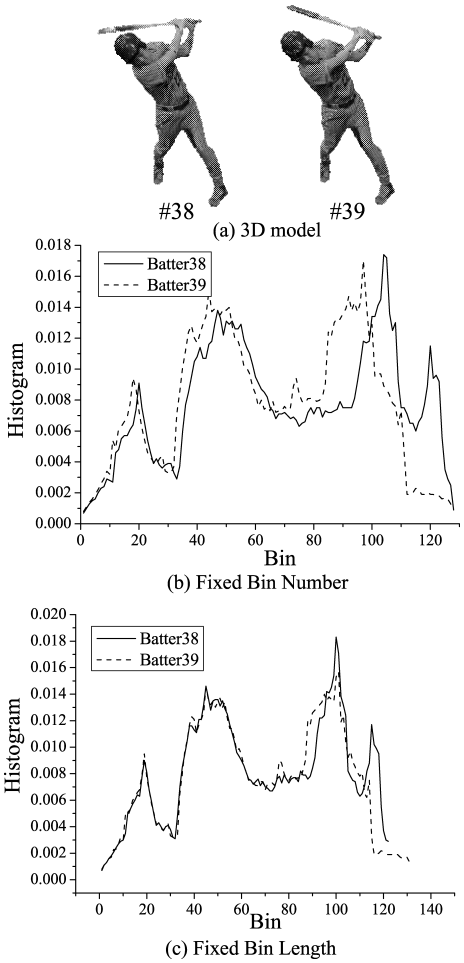

図 6 バッティング・シーケンス 38, 39 フレーム目のヒストグラム（r 成分のみ）：(a) 3 次元モデル；(b) bin 数固定，bin 幅可変（r の最大値・最小値により正規化）；(c) bin 幅固定（0.8），bin 数可変

Fig. 6 Histograms of 38th and 39th frames in batting sequence for r element: (a) 3D model; (b) bin number fixed; (c) bin length fixed.

図 7 にダンスのシーケンスに対して式 (7) を用いてフレーム間の距離を評価した結果を示す．また，比較のために PDH 法[18] による結果もあわせて示す．64 フレーム目から 114 フレーム目までは同一地点に立って回転する動作を行っているのだが，その区間では SCH 法を用いるとフレーム間の距離が大きくなり，何らかの大きな動作が行われていることが検出されている．それに対して，PDH 法ではその区間であまり特徴的な傾向は見られない．これにより，従来の PDH 法に比べて SCH 法の方が 3 次元オブジェクトの動きを表現するのにより適した手法であることが明らかである．

図 8 には bin 幅を変えたときのフレーム間距離の違いを示す．SCH1 と SCH2 を見比べてみるとフレーム間距離に若干の値の違いは見られるものの，傾向などはほとんど同じで，bin 幅を変えても性能には影響がないことが分かる．

ダンス・シーケンスに対する主観的セグメンテーション結果の統計的なばらつきを**図 9** に示す．ただし，ここでは 4 人（50%）以上の投票があった場合を主観評価による正解候補位置と定めており，SCH1 の条件を用いた．アスタリスク（*）が筆者らのシステムによるセグメンテーション結果である．このグラフにより，筆者らのシステムによるセグメンテーション結果が主

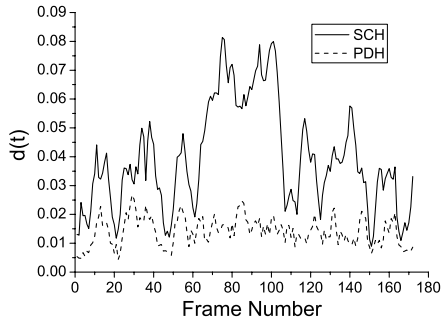

図 7 ダンス・シーケンスに対する SCH 法と PDH 法によるフレーム間距離の違い．SCH 法の方が 64 フレーム目から 114 フレーム目にかけての回転による動きを的確に表現していることが分かる

Fig. 7 Difference in frame-to-frame distance between SCH and PDH. It is demonstrated that SCH can detect the rotating motion from frame #64 to #114 more precisely than PDH.

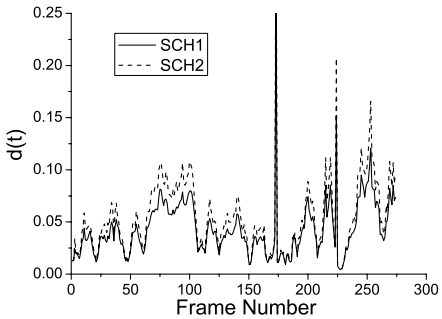

図 8 bin 幅を変えたときのフレーム間距離の違い
Fig. 8 Dependence on bin length.

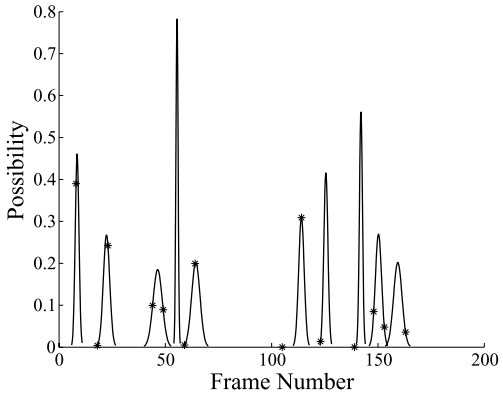

図 9 ダンス・シーケンスに対する主観的投票結果の統計的ばらつき．投票数は 4 に設定
Fig. 9 Possibility distribution of ground truth for Dancer (threshold of valid votes is set as four). Asterisks (*) show the segmentations by SCH1.

観評価結果のばらつきの範囲内にほとんどの場合収まっていることが見てとれる．

図 10 にダンス・シーケンスに対する主観的なセグメンテーション結果と提案手法による結果との比較を示す．これは図 9 に示した結果を各フレームの画像を用いて示したものである．このシーケンスに対しては過検出が 4 つあったものの，検出漏れはなかった．

図 11 に全シーケンスに対するセグメンテーション結果の適合率と再現率を示す．ただし，適合率・再現率とも本論文で提案する統計的手法によって評価した結果である．比較対象として文献 18) で示した PDH 法による結果も示している．グラフ中の数字は何人以上の被験者が投票した場合に正解と見なしたかを表すものである．図 11 中の PDH_{orig} は文献 18) に掲載された実験結果，PDH_{opt} はその後最適化を行い PDH 法で最も性能が良かった実験結果を示している．$Human$ とあるのは 8 人の被験者のうち，明らかに他の被験者とセグメンテーションの仕方の異なる 1 人分

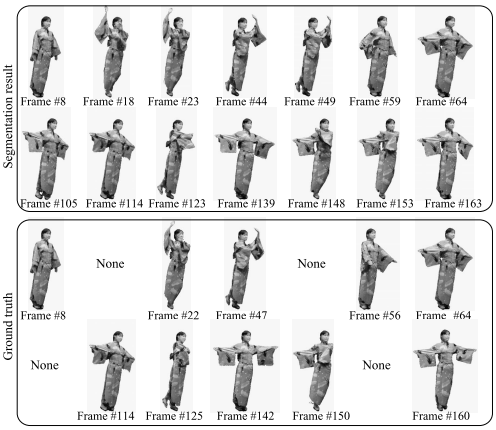

図 10 SCH1 の条件を用いた場合のダンス・シーケンスのセグメンテーション結果．“None” は対応するフレームが存在しなかったことを示す
Fig. 10 Segmentation results. “None” means that there was no corresponding frame.

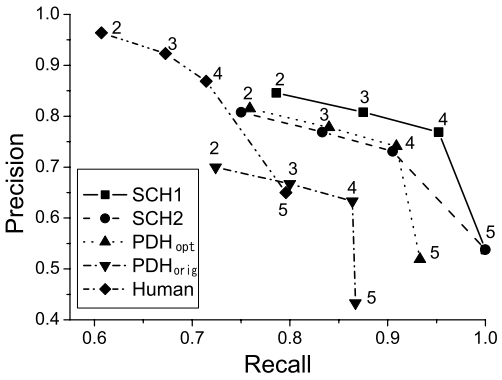

図 11 適合率と再現率．グラフ中の数字は投票数の閾値，すなわち何人以上の被験者が投票した場合を正解位置と判断したかを示す．本論文で開発した SCH 法の性能が PDH 法に比べて大幅に改善されていることが分かる
Fig. 11 Precision and recall rates. The numbers near the points mean the thresholds of valid votes. It can be observed that SCH method developed in this study yields much better performance than PDH.

の結果を取り除いた平均値である．図 11 を見ると分かるとおり，極座標表現を用いた本手法は従来手法に比べて精度の良いセグメンテーションを行えている．SCH 法において bin 幅による性能の違いは無視できる程度である．

セグメンテーション位置をプロットした結果を**図 12** に示す．なお，横軸はフレーム番号，縦軸はフレーム間距離を示している．図 12 を解析したところ，筆者らの提案手法によるセグメンテーションのエラーは以下のような理由により生じていると考えられる．

(1) 人間の目による主観評価においては高次元な

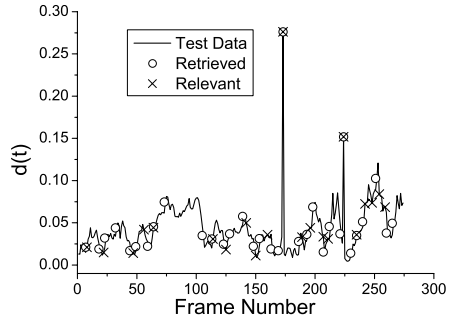

図 12 SCH1 の条件を用いた場合の距離グラフとセグメンテーション位置
Fig. 12 Frame-to-frame distance and segmentation points using SCH1.

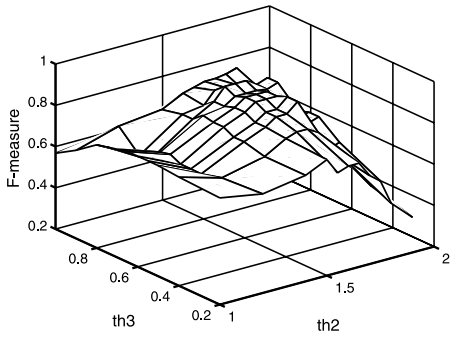

図 13 F 値に対する閾値（$Th2$，$Th3$）の影響
Fig. 13 Dependence of F-measure on threshold values.

動きの意味を理解したうえで行っているが，提案手法では低レベルな信号処理的アプローチをとっているので，おのずと両者のセグメンテーション位置は完全には一致しない．

(2) ポーズ（動きが止まる場面）では人間はセグメンテーションは 1 回起こると判断する場合が多いのに対し，筆者らのシステムでは開始時と終了時の 2 回セグメンテーション位置であると見なしてしまう（図 9，図 10 参照）．

本論文の実験においては式 (8)，(9) で導入した閾値（$Th1$〜$Th3$）は経験的に決定している．そこで，$Th2$，$Th3$ を変化させた場合の F 値の変化の様子を**図 13** に示す．なお，$Th1$ は急峻なシーケンスの切り替わり位置を検出するためのもので，値を変化させてもあまり影響がないので省略してある．今後は動き量や動きの種類に応じた可変的な閾値決定手法の開発も重要である．

6. ま と め

3 次元ビデオという新しい映像表現のデータに対し，動き特徴ベクトルの抽出とそれを用いたセグメンテーションを開発した．頂点群を極座標表現し，そこからヒストグラムを生成することによって，ノイズに強い動き特徴ベクトルの生成が可能となった．本論文で開発した SCH 法は，文献 18) で提案されている PDH 法と比較して，頂点の分布に左右されないヒストグラム生成が可能である．また，複数の被験者による主観評価の結果を統計的に処理し，セグメンテーションの性能を客観的に評価する指標もあわせて開発した．実験の結果，適合率 0.77，再現率 0.95 と文献 18) の手法に比べ精度の良いセグメンテーションが行えることを示した．

謝辞 本論文で使用した 3 次元ビデオ映像は NHK 技研より提供を受けたものである．本研究は一部文部科学省「知的資産の電子的な保存・活用を支援するソフトウェア技術基盤の構築」プロジェクトの支援により行われた．

参 考 文 献

1) Kanade, T., Rander, P. and Narayanan, P.: Virtualized reality: Constructing virtual worlds from real scenes, *IEEE Multimedia*, Vol.4, No.1, pp.34–47 (1997).
2) Zitnick, C.L., Kang, S.B., Uyttendaele, M., Winder, S. and Szeliski, R.: High quality video view interpolation using a layered representation, *Proc. ACM SIGGRAPH 2004*, pp.600–608 (2004).
3) Matsuyama, T., Wu, X., Takai, T. and Wada, T.: Real-time dynamic 3-D object shape reconstruction and high-fidelity texture mapping for 3-D video, *IEEE Trans. Circuit and System for Video Technology*, Vol.14, No.3, pp.357–369 (2004).
4) Tomiyama, K., Orihara, Y., Katayama, M. and Iwadate, Y.: Algorithm for dynamic 3D object generation from multi-viewpoint images, *Proc. SPIE*, Vol.5599, pp.153–161 (2004).
5) Matsuyama, T., Wu, X., Takai, T. and Nobuhara, S.: Real-Time 3D shape reconstruction, dynamic 3D mesh deformation, and high fidelity visualization for 3D video, *International Journal on Computer Vision and Image Understanding*, Vol.96, No.3, pp.393–434 (2004)
6) 中澤篤志，中岡慎一郎，原田貴昭，工藤俊亮，池内克史：視覚による舞踊動作の保存・解析および生成，画像の認識・理解シンポジウム，pp.I-153–I-158 (2002).
7) 池内克史，中澤篤志，小川原光一，高松 淳，工藤俊亮，中岡慎一郎，白鳥貴亮：民族芸能のデジタルアーカイブとロボットによる動作提示，日

本バーチャルリアリティ学会学会誌, Vol.9, No.2, pp.14–20 (2004).
8) 八村広三郎, 中村美奈子：モーションキャプチャデータから舞踊譜 Labanotation の生成, 情報処理学会研究報告, Vol.2001-CVIM-128, pp.103–110 (2001).
9) Idris, F. and Panchanathan, S.: Review of image and video indexing techniques, *Journal of Visual Communication and Image Representation*, Vol.8, No.2, pp.146–166 (1997).
10) Koprinska, I. and Carrato, S.: Temporal video segmentation: A Survey, *Signal Processing: Image Communication*, Vol.16, No.5, pp.477–500 (2001).
11) Rui, Y. and Anandan, P.: Segmenting visual actions based on spatio-temporal motion patterns, *Proc. IEEE Computer Vision and Pattern Recognition*, pp.1111–1118 (2000).
12) Wang, T.S., Shum, H.Y., Xu, Y.Q. and Zheng, N.N.: Un-supervised analysis of human gestures, *Proc. IEEE Pacific Rim Conference on Multimedia*, pp.174–181 (2001).
13) Shiratori, T., Nakazawa, A. and Ikeuchi, K.: Rhythmic motion analysis using motion capture and musical information, *Proc.IEEE Conf. on Multisensor Fusion and Integration for Intelligent Systems*, pp.89–92 (2003).
14) Kahol, K., Tripathi, P. and Panchanathan, S.: Automated gesture segmentation from dance sequences, *Proc. 6th IEEE Int. Conf. on Automatic Face and Gesture Recog.*, pp.883–888 (2004).
15) Barbie, J., Safonova, A., Pan, J.Y. and Faloutsos, C.: Segmenting motion capture data into distinct behaviors, *Proc. Graphics Interface 2004* (*GI'04*), pp.195–194 (2004).
16) Lu, C.M. and Ferrier, N.J.: Repetitive motion analysis: Segmentation and event classification, *IEEE TPAMI*, Vol.26, No.2, pp.258–263 (2004).
17) Takano, W. and Nakamura, Y.: Segmentation of human behavior patterns based on the probabilistic correlation, *Proc. 19th Annual Conf. of the Japanese Society for Artificial Intelligence*, 3F1-01 (2005).
18) Xu, J., Yamasaki, T. and Aizawa, K.: 3D video segmentation using point distance histograms, *IEEE International Conference on Image Processing* (*ICIP*), pp.I-701–I-704 (2005).
19) Xu, J., Yamasaki, T. and Aizawa, K.: Effective 3D video segmentation based on feature vectors using spherical coordinate system, *Meeting on Image Recognition and Understanding* (*MIRU*) *2005*, pp.136–143 (2005).
20) Xu, J., Yamasaki, T. and Aizawa, K.: An evaluation approach for temporal segmentation of 3D videos, *Forum on Information Technology* (*FIT*) *2005*, I-039, pp.91–94 (2005).

(平成 17 年 9 月 21 日受付)
(平成 18 年 3 月 20 日採録)

(担当編集委員 佐藤 真一)

徐 建鋒

2003 年, 中国清華大学にて修士課程修了. 2004 年 10 月より東京大学大学院工学系研究科博士課程在学. 3 次元ビデオのセグメンテーション, 編集に関する研究に従事. IEEE 等会員.

山崎 俊彦

1999 年東京大学工学部電子工学科卒業. 2004 年東京大学大学院工学系研究科電子工学専攻博士課程修了. 2004 年より東京大学大学院新領域創成科学研究科基盤情報学専攻助手. 主として 3 次元ビデオの圧縮や検索を中心とした画像・映像処理に関する研究に従事. IEICE, ITE, IEEE, ACM 等会員.

相澤 清晴（正会員）

1983 年東京大学工学部電子工学科卒業. 1988 年東京大学大学院博士課程修了. 工学博士. 現在, 東京大学大学院新領域創成科学研究科教授. 画像, メディア処理に関する研究に従事. 最近は, ライフログ, 3 次元ビデオ, Web の画像処理等の研究に従事. 日本 IBM 科学賞（2002 年）等受賞多数. IEEE Signal Processing Magazine Editorial Board, IEEE Trans. Multimedia, Trans. CSVT Associate Editor. ACM Multimedia 2005 Short Paper Track co-chair, 第 1 回デジタルコンテンツシンポジウム実行委員長（2005）等多くの学術雑誌, 会議へ参画. IEICE, ITE, IEEE, ACM 等会員.