社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2005－ＳＬＤＭ－121（３）
2005／10／20

信号処理的特徴量による3Dビデオのセグメンテーション

山崎　俊彦　　相澤　清晴

東京大学大学院新領域創成科学研究科
〒277-8561 柏市柏の葉 5-1-5 基盤棟 707
E-mail: {yamasaki, aizawa}@hal.k.u-tokyo.ac.jp

あらまし　近年、複数のカメラを用いて撮影された多視点映像から高精細な動的3次元オブジェクトモデル(3次元ビデオ)を生成する技術の研究が盛んに行われている。今後3次元ビデオのデータベースを構築していくことを考えると、3次元モデルの検索やインデクシング技術の開発が必要不可欠である。これらの実現のためには、まず3次元ビデオのシーケンスを動きの意味の区切れによって細分化(セグメンテーション)する前処理が重要な役割を果たす。しかし、3次元ビデオはモーション・キャプチャによる3次元動きデータ等とは異なり、関節の位置や動き量など構造的な特徴量を抽出することが非常に困難である。そこで本稿ではそのような問題に対し、信号処理的アプローチを用いて3次元ビデオのセグメンテーションを行う手法について検討する。

キーワード　3次元ビデオ、多視点映像、セグメンテーション、シーン分割

Temporal Segmentation of 3D Video Based on Numerical Features

Toshihiko YAMASAKI　and　Kiyoharu AIZAWA

Graduate School of Frontier Sciences, The University of Tokyo

5-1-5 Kashiwanoha, Kashiwa, Chiba, 277-8561 Japan
E-mail: {yamasaki, aizawa}@hal.k.u-tokyo.ac.jp

Abstract　3D video, which is generated from multi-viewpoint images, has been attracting increased 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 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 devision of 3D videos into meaningful and manageable segments. In this paper, we have developed an effective segmentation algorithm using histogram-based feature vector representation. The segmentation algorithm developed in this paper has been applied to three different 3D video sequences, and high recall and precision rates of 0.84 and 0.80, respectively, have been achieved.

Keyword　3D video, segmentation, scene analysis

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盛んに行われている[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。例えば、金出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ている[1]。その後、松山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、冨山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2.5mm～5mm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。即ち、1フレームのデータは3次元モデルの頂点位置・頂点同士の結線情報・各頂点の色(または各3角パッチのテクスチャ)情報の3種類から成っている。また、一般的に

3 次元ビデオのデータはフレーム毎に独立に生成されるため、例え隣接するフレーム間同士であっても頂点数や結線情報は異なることが多いのが特徴である[1]。

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、取得ばかりでなく圧縮や検索・編集技術の開発が必要不可欠である。しかし、上記に述べた通り 3 次元ビデオはフレーム間で頂点数や結線情報が保存されないために解決すべき課題が多い。

3 次元モデルの時間的・動的変化を考慮に入れた圧縮技術に関しては波部ら[6]や韓ら[7]、Muller ら[8]の研究例が報告されている。波部らは skin-off と呼ばれる 3 次元メッシュモデルを 2 次元平面に展開する技術を用いて従来の 2 次元動画像圧縮技術を適用するアプローチをとっている[6]。Muller らは頂点の数や結線情報が隣接フレーム間で変化しないという限定条件があるが、2 次元動画像圧縮に広く用いられているブロックマッチング法を 3 次元に拡張した方式を提案している[8]。また、韓らは Muller らに先立ち、頂点数や結線情報が保存されない一般的な 3 次元ビデオ・データにも適用可能な拡張ブロックマッチング法を開発し、既に良好な圧縮結果を報告している[7]。

以上に述べたように圧縮に関してはいくつかの研究例が報告されている一方で、3 次元ビデオの検索や編集に関しては未だ報告例がない。3 次元ビデオはデータ量が膨大であるため[2]、検索や編集を可能とするためにはそれらの技術の開発に先立ってモデルの時間的変化を考慮したシーケンス分割、すなわちセグメンテーション技術の開発が重要である。これによって 3 次元ビデオを意味のある単位で分割することができ、その後の処理を容易に行えるようになる。

これまで 2 次元映像のセグメンテーションについては多数の研究例が報告されているが[9][10]、それらは主に映像の 2 次元色情報の統計的処理に基づくものである。そのためポリゴンの頂点座標や結線情報が主な要素である 3 次元ビデオ・データにそのまま拡張するのはふさわしくない。筆者らの研究グループでは 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頂点の距離や角度の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[11][12]。我々の知る限り 3 次元ビデオ・セグメンテーションの研究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[11][12]。また、文献[13]では特に人間の動きに着目し、動きの意味の区切れにおいてはオブジェクトを囲むバウディング・ボックスの体積が時間的に極小値になるというセグメンテーションモデルを提唱し、適合率 95%、再現率 76%と良好な結果を得た。しかし、このモデルに適合しない例外的な動作が多く一般的なセグメンテーションには不向きであるという問題があった。

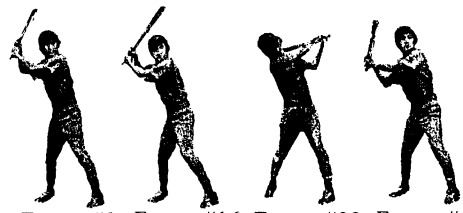


図 1. 3 次元ビデオの例。ここでは一視点からの画像のみを示す。

本論文の目的は、3 次元ビデオからよりロバストな形状・動き特徴ベクトルを抽出し、高精度なセグメンテーション技術を開発することである。本論文では、静止 3 次元オブジェクトの形状特徴を表現する優れた手法の 1 つとして提案されている Shape Distribution 法[14]を 3 次元ビデオ・データに最適化し、フレーム毎に生成されたヒストグラムの時間変化を解析することによりセグメンテーションを行った。その結果、適合率 0.84、再現率 0.80 という今までに我々が提案してきた手法と比べて遜色のない結果が得られ、3 次元ビデオの検索に向けた新しい可能性を示すことが出来た。

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・データは冨山ら[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。

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、及び各頂点の色の 3 種類のデータから構成されている(図 2 参

[1]松山らは数フレームに渡り頂点数や結線情報を保ったまま生成する技術の開発も行っている[5]。

[2]例えば、冨山らの例では 1 フレームあたり 5～10 万個の頂点と同数の色情報、10～20 万個の結線情報のデータがあり、VRML で記述した場合 5MB～10MB/frame になる。

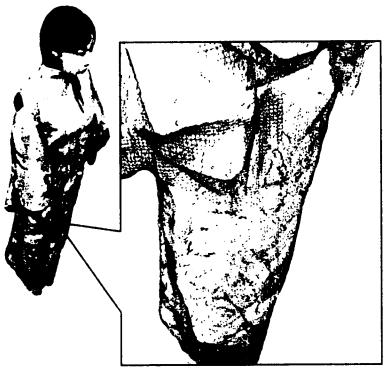


図 2. 3 次元モデルの拡大図。

表 1. 用いた 3 次元ビデオのデータ諸元。

シーケンス	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

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。それぞれの 3 次元ビデオを一定方向から見た映像を[15]-[17]に示す。3 次元ビデオは多視点映像から各フレーム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間であっても異なるのが特徴である。

3. セグメンテーションのアルゴリズム

3.1. システムの概要

本論文では 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。例えば、図 1 に示したバッターのシーケンスを例に取ると「構え」、「ヒッティング」などに分割できると考えられる。ここで注意しなければならないのが、どんな動きに対してもセグメンテーションが行えるよう、予備的な動きの分類・定義などは行っていないということである。そのため、本論文ではオブジェクトの動き、即ち形状・姿勢の時間変化をロバストに抽出できるような特徴ベクトルを用いた信号処理的なアプローチを採用する。提案手法では、ヒストグラム・ベースの特徴ベクトルを採用した。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。生成されたヒストグラムを用い、3.3 章で述べるアルゴリズムに基づいてセグメンテーション位置を判断する。提案手法による実験結果は第 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

本論文では 1 フレーム毎にオブジェクトの形状特徴ベクトルを抽出し、その時間変化を解析することでセグメンテーションを行う。静止 3 次元オブジェクトから形状特徴を抽出する手法は 3 次元オブジェクト検索技術のために多くの手法が検討されている[14][18][19]。例えば Hilaga らはオブジェクトから Reeb Graph と呼ばれるグラフ構造によって形状の特徴を表現し、位相関係のマッチングによって類似度を評価している[18]。また、Chen らはオブジェクトを複数の視点から見たときのシルエット画像を静止し、それらをオブジェクト間で比較することで類似度を評価している[19]。しかし、これらは例えばコーヒーカップと飛行機の 3 次元オブジェクトモデルのように全く形状が異なる物同士が似ていないということは判断できても、3 次元ビデオのフレーム間の微妙な形状・動きの変化を表現するのには適していない。

様々な静止オブジェクト検索のための特徴抽出アルゴリズムの中で 3 次元ビデオの動き解析応用に適していると思われるのが Osada らによって開発された Shape Distribution 法[14]である。この手法は 3 次元オブジェクトの表面にランダムにばらまかれた点同士のユークリッド距離を計算し、そのヒストグラムによってオブジェクトの形状特徴を表現するアルゴリズムである。この手法では頂点の数を N としたとき、計算量が $O(N^2)$となる問題があるが、オブジェクト表面の点をランダムサンプリングすることによって計算量の爆発を防いでいる。

しかし、オリジナルの Shape Distribution 法を用いると、ランダムサンプリングで頂点を選択するため同一のフレームに複数回試行すると生成されるヒストグラムに不確定的なノイズが発生する。また、隣り合うフレーム間におけるオブジェクトの形状変化は微妙なものであるため、ヒストグラムにそのノイズの影響を受けやすい。そこで本論文では各フレームで安定してヒストグラムが生成されるよう、方式の改善を行った。提案手法ではまず頂点群のクラスタリングを行い、1024 点の代表点を抽出する。クラスタリングには頂点のユークリッド座標(x, y, z)の 3 次元を 1 つのベクトルとしたベクトル量子化を用いる。これによって図 2 に示すように選択される代表点がオブジェクトの表面にほぼ一様に分布させることができ、安定したヒストグラムの生成が可能となる。ヒストグラム生成時には距離の最大値・最小値の間を等しく 1024 の区間に分割し

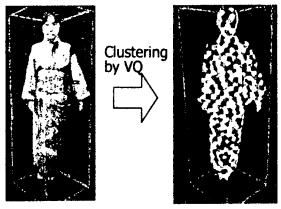


図 2. ベクトル量子化を用いた 3 次元オブジェクト頂点のクラスタリング結果。

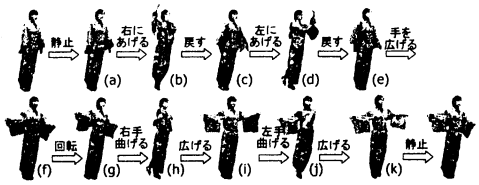


図 3. 被験者 8 人によるダンス映像のセグメンテーション結果。

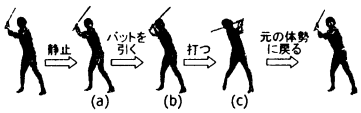


図 4. 被験者 8 人によるバッティング映像のセグメンテーション結果。

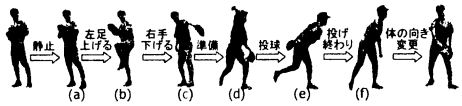


図 5. 被験者 8 人によるピッチング映像のセグメンテーション結果。

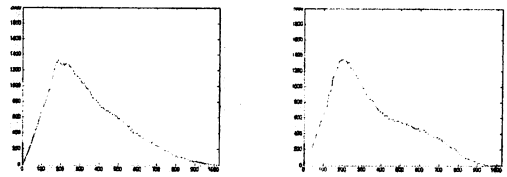
た。この代表点数 1024、ビン数 1024 という条件は文献[14]で精度と計算コストのトレードオフから最適であるとされていたものである。今後 3 次元ビデオの形状・動き特徴を表現するのに最適な条件を検討するのも今後の重要な課題の 1 つである。

3.3. セグメンテーション位置検出

動きの意味の区切れは一般的に動きが止まったとき、または動きの方向・種類が変わったときに生じると考えられる。動きの方向や種類が変わるときも、そのために一時的に動きの変化量は小さくなる。よって、本論文では隣り合うフレーム間の特徴ベクトルの距離を計算し、変化が極小値になったときをセグメンテーションの位置とする。また、「静止」という動作に対応するため、変化が一定だったときの両端もセグメンテーション位置とした。ただし、フレーム間距離の計算結果に対しては前後 2 フレームの距離の相加平均を取り、時間方向に平滑化を施す。

0 フレーム目　　22 フレーム目

(a)



0 フレーム目　　22 フレーム目

(b)

図 6. 生成された特徴ベクトル例: (a) 3 次元オブジェクトモデル; (b) ヒストグラム。

4. 評価手法

シーン分割の正解位置については、客観的な区切り位置というのは存在しない。そこで 3 次元ビデオのシーン分割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、主観評価によって正解位置を定めた。本実験では、8 人中半数の 4 人以上が区切り位置であるとしたフレームをシーン分割の正解位置とした。ただし、オブジェクトが動いていることを考慮し、±3 フレームのばらつきに関しては同一の区切り位置であると見なした。それぞれの 3 次元ビデオに対するシーン分割の結果を図 3〜図 5 に示す。但し、始まりと終わりのフレームは区切りとして自明なので、ここでは対象としない。

5. 実験結果と考察

図 6 に生成された特徴ベクトルの代表例を示す。0 フレーム目では着物を着た女性が両手を下げて静止しており、22 フレーム目では両手を右に挙げて体をねじっている。図 6(a)に示したオブジェクトの形状変化に合わせて、図 6(b)のヒストグラムも微妙に変化していることが見て取れる。ヒストグラムが余り大きく変化しないのは、胴体や足の部分が両者で殆ど同じ位置関係にあるためである。

0 フレーム目のモデルに対しランダムサンプリングによって代表点を選択した場合の、生成されるヒストグラムのばらつきを図 7 に示す。同一のモデルに対してヒストグラムを生成しているのにも関わらず、その形状には最大値の差や山型形状の傾きの差など様々な

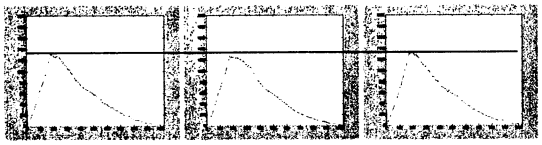


図 7. 0 フレーム目のモデルに対しランダムサンプリングによって代表点を選択した場合の、生成されるヒストグラムのばらつき。

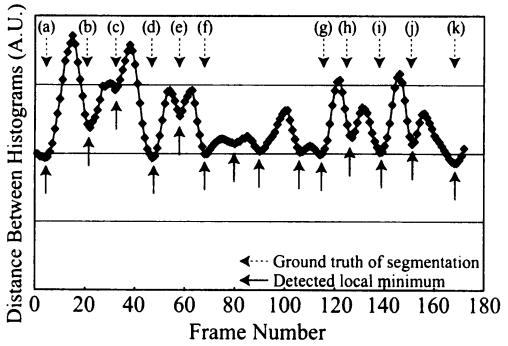


図 8. ダンス・シーケンスに対する 3D ビデオのセグメンテーション結果。点線矢印の位置が 8 人の主観評価によるセグメンテーション位置、実線矢印が提案アルゴリズムにより得られたセグメンテーション位置。

ばらつきが生じていることがわかる。それに対し、提案手法では代表点がオブジェクト表面にほぼ一様に分散されるため似た形状のオブジェクトに対しては似た形状のヒストグラムが生成されることを確認した。

図 8 に 8 人の主観評価によるセグメンテーション結果と提案手法による実験結果を示す。図 8 に示す通り、フレーム間動きの変化量が極小値になる時点と主観評価によるセグメンテーション位置はよく一致していることがわかる。しかし、80、90、106 フレーム目に過検出が起きている。68 フレーム目から 114 フレーム目までは両手を広げて左回りに回転するシーンである。これらの時点による映像を解析したところ、軸足を変えているときであることがわかった。つまり、軸足を変える際回転の速度が低下し、結果としてヒストグラム間距離に極小値が生じていた。「軸足の切り替え」という意味ではセグメンテーション位置であるといえなくもないが、人間の目による主観評価においては軸足の切り替えよりもより大きな動きである「両手を広げて回転」の方に注意が向けられるために正解位置としては定義されていない。

このように、人間の目による主観評価においては高次元な動きの意味を理解した上で行っているが、提案手法では低レベルな信号処理的アプローチを取っているので自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。

表 2. ダンス・シーケンスに対するセグメンテーション結果。

Frame	ID	motion	N. B.
0		静止	
5	a	両手を右上に挙げる	
22	b	両手を戻す	
32	c	両手を左上に挙げる	
47	d	両手を戻す	
58	e	両手を伸ばす	
68	f	回る	
114	g	止まって右手を左側に曲げる	
127	h	右手を戻す	
138	i	左手を右側に曲げる	
151	j	左手を戻す	
168	k	静止	
172		終了	

* #80, #90, #106: 過検出

表 2〜表 4 に各シーケンスに対するセグメンテーション結果の詳細を示す。ダンス・シーケンスに対しては検出漏れは 1 件もなく、過検出が 3 件あった。また、バッティング、ピッチングのシーケンスに関しては過検出は 0 件、検出漏れが数件みられた。特に、ピッチング動作のように一連の動作でありながら途中で動きの意味が変わる動作の検出は難しいことが実験の結果わかった。全体の精度としては適合率 0.84、再現率 0.80 であった。

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、動き特徴量の抽出とそれを用いたシーン分割手法を開発した。静止 3 次元オブジェクト検索に用いられる Shape Distribution 法を 3 次元ビデオに最適化し、安定した特徴ベクトルの抽出が可能となった。各フレームにおいて抽出された特徴ベクトル同士の距離を計算し、動きの変化が小さくなった時点をセグメンテーション位置とすることにより適合率 0.80、再現率 0.84 の精度を得た。

3 次元ビデオは取得・生成自体がまだ新しい技術であるため、本論文で用いることのできたデータの量には限りがあった。今後はさらにテスト・データを増やして提案手法の妥当性を検討していく必要がある。

また、第 5 章で考察したように主観評価は動きの高次的な理解に基づいて行われるのに対し、提案手法は信号処理的なローレベルな解析によって行われている。

表 3. バッティング・シーケンスに対するセグメンテーション結果。

Frame	ID	motion	N. B.
0		静止	
16	a	バットを引く	miss
23	b	打つ	
34	c	静止	
50		終了	

* 過検出なし

表 4. ピッチング・シーケンスに対するセグメンテーション結果。

Frame	ID	motion	N. B.
0		静止	
11	a	右手を胸元に	miss
18	b	右手下げる	miss
24	c	準備	
30	d	投球	miss
35	e	投げ終わり	
45	f	体の向き変更	
50		終了	

* 過検出なし

より高精度なセグメンテーションを実現するためには統計処理に基づく手法等が必要であり、現在検討中である。

7. 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基盤技術の構築」プロジェクトの支援により行われた。

文　献

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

[2] C.L. Zitnick, S.B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High quality video view interpolation using a layered representation," Proceedings of ACM SIGGRAPH 2004, pp. 600-608, 2004.

[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 Trans. Circuit and System for Video Technology, vol. 14, no. 3, 2004.

[4] K. Tomiyama, Y. Orihara, M. Katayama, and Y. Iwadate, "Algorithm for dynamic 3D object generation from multi-viewpoint images," Proceeding of SPIE, vol. 5599, pp. 153-161, 2004.

[5] 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," International Journal on Computer Vision and Image Understanding, vol. 96, no. 3, pp. 393-434, 2004.

[6] H. Habe, Y. Katsura, and T. Matsuyama, "Skin-off: Representation and Compression Scheme for 3D Video," Picture Coding Symposium (PCS), San Francisco, 2004.

[7] S. Han, T. Yamasaki, and K. Aizawa, "Compression of 3D Video Using 3D Block Matching Algorithm," Technical Report of IEICE, IE 2005-20, vol. 105, no. 161, pp. 13-18, 2005 (in Japanese).

[8] K. Muller, A. Smolic, M. Kautzner, P. Eisert, and T. Wiegand, "Predictive Compression of Dynamic 3D Meshes," The 2005 IEEE International Conference on Image Processing (ICIP2005), pp. I-621-I-624, 2005.

[9] I. Koprinska, and S. Carrato, "Temporal video segmentation: A Survey," Signal Processing: Image Communication, vol. 16, no. 5, pp. 477-500, 2001.

[10] Y. Aslandogan and C. Yu, "Techniques and Systems for Image and Video Retrieval," IEEE Transactions on Knowledge and Data Engineering, vol. 11, no. 1, pp. 56-63, 1999.

[11] J. Xu, T. Yamasaki, and K. Aizawa, "3D Video Segmentation Using Point Distance Histograms," The 2005 IEEE International Conference on Image Processing (ICIP2005), pp. I-701-I-704, 2005.

[12] J. Xu, T. Yamasaki, and K. Aizawa, "Effective 3D Video Segmentation Based on Feature Vectors Using Spherical Coordinate System, Meeting on Image Recognition and Understanding (MIRU) 2005, pp. 136-143, 2005.

[13] 山崎俊彦, 徐建鋒, 相澤清晴, 「ポリゴン頂点の主成分分析による 3D ビデオの動き特徴量抽出とシーン分割」, 情報科学技術フォーラム, FIT2005, I-040, pp. 95-98, Tokyo, Sep.7-9, 2005.

[14] R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin, " Shape Distributions," ACM Transactions on Graphics, 21(4), pp. 807-832, October 2002.

[15] www.hal.k.u-tokyo.ac.jp/~yamasaki/IE2005_10/nihon-buyou.gif.

[16] www.hal.k.u-tokyo.ac.jp/~yamasaki/IE2005_10/batting.gif.

[17] www.hal.k.u-tokyo.ac.jp/~yamasaki/IE2005_10/pitching.gif.

[18] M. Hilaga, Y. Shinagawa, T. Kohmura, and T. L. Kunii, "Topology matching for fully automatic similarity estimation of 3d shapes," In Proceedings of ACM SIGGRPAH 2001, pp. 203-212, , 2001.

[19] D.Y. Chen, M. Ouhyoung, X.P. Tian, Y.T. Shen, and M. Ouhyoung, "On visual similarity based 3D model retrieval," In Proc. Eurographics, 2003.