I-040

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

Motion Feature Extraction and Temporal Segmentation of 3D Videos Based on Principal Component Analysis of Polygon Vertices

山崎 俊彦† **徐 建鋒‡** **相澤清晴†**
Toshihiko Yamasaki Jianfeng Xu Kiyoharu Aizawa

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動的 3D オブジェクトモデル(以後「3D ビデオ」と呼ぶ)を生成する研究が盛んに行われている[1]-[4]。これは従来の CG による 3D 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、時間変化を追うことができるために新しい映像表現として注目を浴びている。

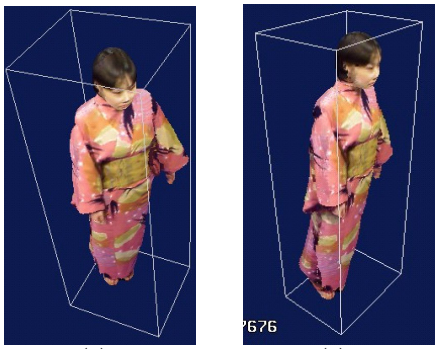
3D ビデオは新しい研究分野であるため、データの取得についてもまだ取り組むべき課題が多い。さらに 3 次元ビデオをアーカイブ化し実用的に利活用できるようにするためには、その圧縮や検索技術が必要となる。後者の検索のためには、3D ビデオの時間的変化を考慮したシーン分割技術の開発が重要である。シーン分割は検索の前処理としてばかりでなく、編集の助けにもなる。

これまで 2 次元映像のシーン分割については多数の研究例が報告されているが[5]、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 3D ビデオデータにそのまま拡張するのはふさわしくない。我々の研究グループでは 3D ビデオのシーン分割について取り組み、これまでに基準点と 3 次元頂点の距離ヒストグラムを用いたシーン分割を提案してきた[6][7]。我々の知る限り 3D ビデオのシーン分割の研究としては初めての試みである。この手法では精度よくシーン分割を行うことができる反面、距離ヒストグラムと動きの大きさや回転といった動き特徴量との対応をとるのが困難であるため、類似動作の検索などには応用できない問題があった。さらに、分割点を決定する最適閾値の決定も問題の 1 つとなっていた。

そこで、本論文では特に人物の動きをセグメンテーションすることを対象とし、3D オブジェクトを囲む直方体の枠(バウンディングボックス)の形状変化に注目して動き特徴量を表現する手法を提案する。同じ動きであってもオブジェクトの向き・姿勢によってバウンディングボックスの形状は変化するため、頂点座標を主成分分析して姿勢補正を行う手法を開発した。3D オブジェクトに対し主成分分析を行う研究としては、類似形状オブジェクトの検索の際に視点依存性をキャンセルするために利用した例がある[8]。しかし、文献[8]では回転角度の利用や 3D オブジェクトの時間的変化は一切扱っていない点で本論文とは異なる。提案手法により、3D オブジェクトの空間的移動・回転・動きの大きさ・類似動作の繰り返しなどの動き特徴が表現可能となった。また、バウンディングボックスの体積変化の極小位置を利用してシーン分割を行う手法も併せて提案する。

†東京大学大学院新領域創成科学研究科

‡東京大学大学院工学系研究科



(a) (b)

図 1. 3 次元オブジェクトのバウンディングボックス: (a) 元データ; (b) 主成分分析による姿勢補正後。

２．主成分分析によるオブジェクト姿勢補正

本論文で使用する 3D ビデオ・データは 1 フレームずつ独立に生成され、それぞれのフレームではポリゴンの頂点・結線情報・各頂点の色の 3 種類のデータより成っている。また、汎用性のために 3D モデル記述の標準言語である VRML によって記述されている[3]。

3D オブジェクト座標の x, y, z 各成分の最大値・最小値を検索し、オブジェクトを囲む直方体すなわちバウンディングボックスを考える。本論文では 3D オブジェクトの動きに合わせて変化するバウンディングボックスの時間的変化を用いて動き特徴量を抽出する。その際問題になるのが、同じ動きであってもオブジェクトの向き・姿勢によってバウンディングボックスの形状が変化する点である(図 1 参照)。この問題に対処するため、頂点座標に対して主成分分析を施す。主成分分析とは多次元ベクトルの要素間の相関を無くす手法のことで、より低次元の要素数で元のベクトルの特性を記述するための多変量データ解析手法として頻繁に用いられている[9]。しかし、本論文では主成分分析をベクトルの次元数を減らす為ではなく、3D オブジェクトの頂点分布の偏りを利用してオブジェクトの回転を補正するために用いる。例えば人体は前後方向よりも左右方向の方が頂点の分布幅が大きい。この特徴を利用して、体の正面が常に座標系の y 軸と垂直になるようにオブジェクトの姿勢を補正することができる(図 1(b)参照)。この主成分分析による姿勢補正は、人工的な球体や立方体など対称なオブジェクトに対しては有

効でない。しかし、本論文で対象としているのは実写 3D オブジェクトなので完全に対称な物体は存在しないものして議論を進める。

本論文では主に人体の動きを対象としているために 3D オブジェクトは xy 平面に対してほぼ垂直であると仮定する。これにより 3 次元座標のうち x, y 成分のみに対して主成分分析を行えばよいので計算量が軽減できる。またオブジェクトの回転も 2 次元的に扱うことができる。以上のことをふまえた主成分分析のアルゴリズムを以下に示す。

1. ポリゴンの頂点座標群 $P=\{\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_i, \cdots \boldsymbol{p}_n\}$, $\boldsymbol{p}_i=(p_{i1}, p_{i2}, p_{i3})$は、予め x, y 成分の重心が座標の原点にくるように平行移動してあるものとする。
2. x, y 成分の分散共分散行列 M を計算する。

$$M = \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} p_{i1}^{\,2} & p_{i1}\cdot p_{i2} \\ p_{i1}\cdot p_{i2} & p_{i2}^{\,2} \end{pmatrix}$$

3. 行列 M の固有ベクトル行列 Me を計算する。ただし、Me は固有値の大きい順に並んでいるとする。
4. 姿勢補正後の座標 $\boldsymbol{p}_i^*$ を $\boldsymbol{p}_i^*=Me*\boldsymbol{p}_i$ により計算する。

ここで、主成分分析は単純に頂点の分布を解析しているだけなので、例えばオブジェクトの前面と後面といった形状の意味を理解した上での姿勢補正ができない。そのため頂点の分布によっては、時間的に連続するフレーム間で、補正された 3D オブジェクトの向きが 180 度異なる場合がある。一般に 1 フレーム後に 90 度以上の回転をするような大きな動きはないと考えられるので、そのような異常な回転角度が検出された場合はさらに 180 度回転の補正を行う。

3．動き特徴量を用いたシーン分割

第 2 章で述べた主成分分析によって、人物の体を対象とした場合おおよそ左右方向が x 軸、前後方向が y 軸、上下方向が z 軸に対して平行になる(図 1(b)参照)。そのため、人体の動きに対してバウンディングボックスの x, y, z 各成分の幅の変化を追跡することで体の横、縦、高さ方向の広がりを表現することができる。さらに、バウンディングボックスの幅の時間微分によって動きの大きさも表現できる。これらの他にも 3D オブジェクトの重心位置を追跡することでオブジェクトの平行移動を、主成分分析の結果得られた中でもっとも大きな固有値に対応する固有ベクトルと x 軸とのなす角によってオブジェクトの回転角が表現できる。ただし、回転角に関しては第 2 章に述べたとおり補正がなされているものとする。

また、類似な動作をしている場合、バウンディングボックスの形状も似たような変化がおきるので、この特徴を利用したパターン認識によって類似動作の検索を行うことが可能である。

動きの意味の区切れに関しては複数の 3D ビデオを解析した結果、手を左右や上下に伸ばす、または折りたたむなどバウンディングボックスの x, y, z いずれかの幅が極小値になる場合が多いことがわかった。また、一連の動作の過程で生じる動作の区切り目と関係のない極小値については、他の成分の幅が逆に大きくなることも明らかになった。以上の特徴を利用して、バウンディングボックスの体積を動きの意味の特徴量として用いる。すなわち、バウンディングボックスの体積が極小値となるフレームをシーン分割の位置として扱う。ただし、人間の動作では特に歩行の時に足や手が前後に動くために、単純なバウンディングボックスでは手足の動きに合わせて体積が微妙に変化してしまう。これがシーンの過分割を生じる原因となる。しかし、歩行などは大きな動きを伴わないので、バウンディングボックスの体積を変化させる原因、すなわちバウンディングボックスの境界付近に存在する頂点の数は非常に少ない。この特徴を利用して、バウンディングボックスの両端から全頂点数の 5%に当たる頂点を無視してバウンディングボックスを定義することで過度な極小値の検出を防止する。さらに、無視することのできなかった極小値やノイズ的に発生する極小値を除去するために、極小値の両端にある極大値の大きさとの差が 5%未満となる場合、その極小値は無視する。

図 2. 被験者 8 人による日本舞踊映像のシーン分割結果。

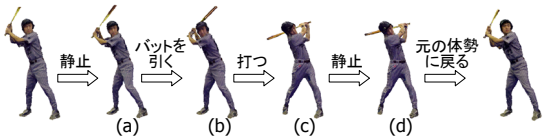


図 3. 被験者 8 人によるバッティング映像のシーン分割結果。

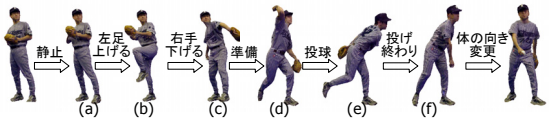


図 4. 被験者 8 人によるピッチング映像のシーン分割結果。

4．実験と考察

実験には NHK 技研より提供を受けた日本舞踊(全 173 フレーム)、バッティング(全 51 フレーム)、ピッチング(全 51 フレーム)3 種類の映像を用いた[3]。また、フレームレートは取得環境の仕様から 10fps である。それぞれの 3D ビデオを一定方向から見た映像を[10]-[12]に示す。

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 3D 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 3D 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 3D ビデオに対するシーン分割の結果を図 2～図 4 に示す。但し、始まりと終わりのフレームは区切りとし

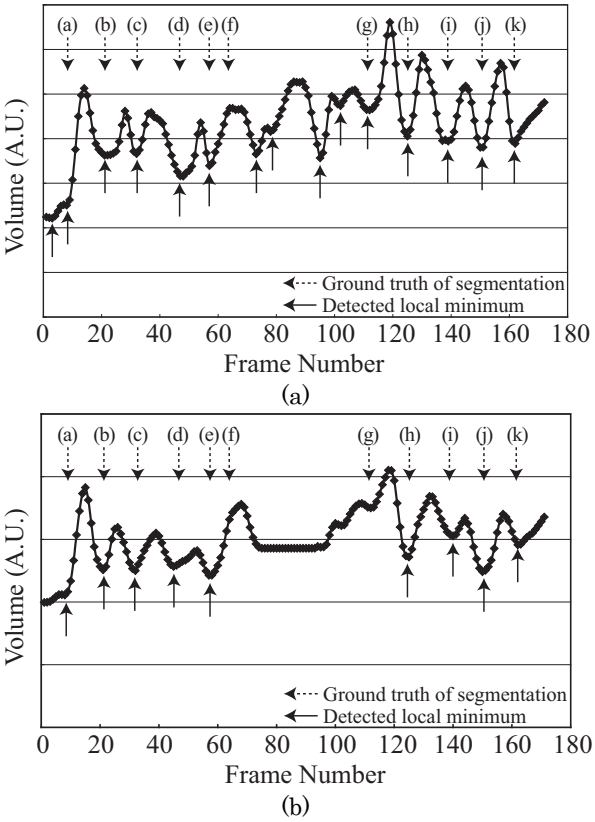


図 5. バウンディングボックスの体積の極小値を用いた 3D ビデオシーン分割結果: (a) ノイズ除去処理前; (b) ノイズ除去処理後。図中(a)〜(k)は図 2 に示した、シーン分割正解位置。実線矢印で示した位置が、提案アルゴリズムにより得られたシーン分割位置。

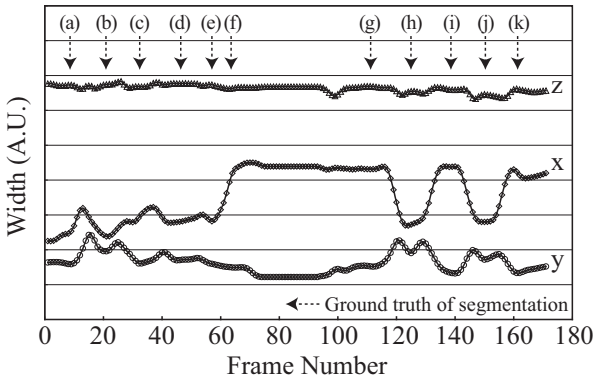


図 6. バウンディングボックスの x, y, z 方向の幅。

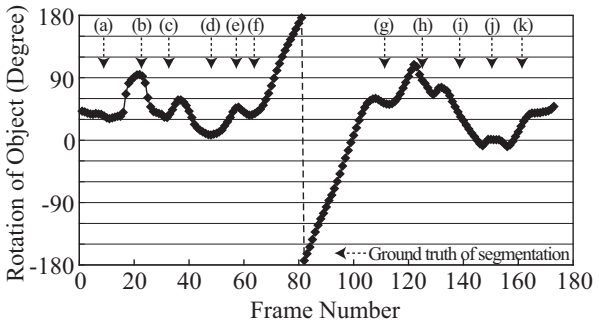


図 7. 主成分分析により得られたオブジェクトの回転角度。

て自明なので、ここでは対象としない。これ以降、特に断りがない限り日本舞踊の 3D ビデオを例に取り実験結果を示す。

3D オブジェクトに対して主成分分析を施し、姿勢補正を行った結果を図 1(b)に示す。バウンディングボックスが体の縦・横・高さに対して平行になり、正しく姿勢補正ができていることがわかる。また、今回実験に使用した 3D ビデオの全フレームに対して正しく姿勢補正ができていることを確認した。

バウンディングボックスの体積変化とシーン分割の正解位置並びに提案アルゴリズムによるシーン分割位置を図 5 に示す。点線矢印が正解位置、実線矢印が提案アルゴリズムによる分割位置である。図 5(a)と図 5(b)を見比べるとわかるとおり、提案するノイズ除去手法によって過分割が大幅に減っていることがわかる。その代わりに、検出漏れが(f)1 つだけだったのが (f), (g)の 2 つとなっている。ただし、これはもともと(g)が歩行動作時の足の影響でたまたま検出されていたものだと考えられる。提案手法ではバウンディングボックスの体積変化の極小値を用いてシーン分割を行うので、体を伸ばすなどバウンディングボックスの体積を広げる動作をした直後に別の動作に移行すると、その部分が検出できない。位置(g)についてはオブジェクトが回転するのをやめたのみで、バウンディングボックスの形状は変化していないと考える方が正しいので、提案手法では本来検出されない区切り位置である。この問題については今後別の特徴量を導入する必要がある。

x, y, z それぞれの幅をプロットしたものを図 6 に示す。シーン分割位置では x, y, z いずれかの成分が極小値になっている。ただし、第 3 章で議論したとおりすべての極小値の位置がシーン分割位置となっているわけではない。また、動きの大きさに対応して幅も変化し、(g)〜(h)と(i)〜(j)のように似た動作が起こっているときにはバウンディングボックスの幅のパターンも類似していることがわかる。今回は幅のみを考慮しているため、左右/上下対称な動きは同じパターンとなって現れる。これによってバウンディングボックス幅のパターン解析という比較的簡単なアプローチによって類似動作の検索等が行える可能性が示せた。

図 7 に 3D オブジェクトの回転検出の結果を示す。第 2 章に述べた回転補正の結果、オブジェクトの回転が正しく検出できていることがわかる。特に右手・左手を頭上に挙げるときの体のひねり(図 2 中(a)〜(d)の区間)や、両手を左右に広げて一定速度で回転している様子(図 2 中(f)〜(g)の区間)がグラフから詳細に見て取れる。フレーム#80 で角度の不連続が起きているように見えるが、これは角度を-180 度〜+180 で表現しているためであり、実際は連続している。オブジェクトの回転角に対しても、シーン分割位置のほとんどが極小値及び極大値のいずれかの極値位置に対応していることがわかる。

3 種類の 3D ビデオに対するシーン分割結果を表 1〜表 3 に示す。すべての 3D ビデオにシーン分割を施した結果、正解位置 21 個に対し、過検出 1、検出漏れ 5 となり、適合率は 95%、再現率は 76%であった。距離ヒストグラム

表 1. 日本舞踊映像のシーン分割結果

Frame	ID	motion	N.B.
0		静止	
8	a	両手を右上に挙げる	
22	b	両手を戻す	
33	c	両手を左上に挙げる	
47	d	両手を戻す	
56	e	両手を伸ばす	
64	f	回る	miss
114	g	止まって右手を左側に曲げる	miss
125	h	右手を戻す	
142	i	左手を右側に曲げる	
150	j	左手を戻す	
160	k	静止	
172		終了	

* 過検出なし

によるシーン分割[6][7]に比べて、特に過検出が少ない。過検出はバッティングの 3D ビデオでバットが人体に対して垂直になったときに偶然体積の極小値が生じたものである。バットの動きは人体の動きに比べてより大きな影響を及ぼす。逆に検出漏れは若干多くなった。特にピッチング動作では投球モーションのときなど体を伸ばした後すぐに次の動作に移るようなシーンが多いため、体積の極小値による提案手法では検出漏れが多く発生していると考えられる。

5．まとめ

3D 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。ポリゴン頂点の主成分分析によって 3D オブジェクトの姿勢を補正し、その後バウンディングボックスの幅や体積の変化を解析することでオブジェクトの空間的移動・回転・動きの大きさ・類似動作の繰り返しなどの動き特徴が表現可能となった。また、バウンディングボックスの体積変化の極小値探索という平易な方法でロバストにシーン分割を行う手法を提案した。

本論文では取得できたデータの量に限りがあったため、今後さらにテスト・データを増やして検討を行っていく必要がある。提案手法では体を伸ばしてすぐに次の動きに移行するような動作の検出が不得手であるため、今後距離ヒストグラムによるシーン分割など別のアプローチとの融合を試みる。さらに、類似動作の検索についてもその可能性を示したのみであるので、今後さらに検討を行う予定である。

6．謝辞

本論文で使用した 3D 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。

表 2. バッティング映像のシーン分割結果

Frame	ID	motion	N.B.
0		静止	
16	a	バットを引く	
23	b	打つ	
34	c	静止	
38	d	元の体勢に戻る	
50		終了	

* #45: 過検出

表 3. ピッチング映像のシーン分割結果

Frame	ID	motion	N.B.
0		静止	
11	a	右手を胸元に	miss
18	b	右手下げる	
24	c	準備	miss
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

7．参考文献

[1] S. Moezzi, L. Tai, and P. Gerard, “Virtual view generation for 3-D digital video,” IEEE Multimedia, pp. 18–26, 1997.
[2] G. Cheung and T. Kanade, “A real time system for robust 3D voxel reconstruction of human motions,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 714–720, 2000.
[3] Y. Iwadate, M. Katayama, K. Tomiyama, and H. Imaizumi, “VRML animation from multi-view images,” Proc. IEEE ICME 2002, vol. 1, pp. 881-884, 2002.
[4] 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. on Circuits and Systems for Video Technology, Vol.CSVT-14, No.3, pp.357-369, 2004.
[5] I. Koprinska, and S. Carrato: “Temporal video segmentation: A Survey”, Signal Processing: Image Communication, Vol. 16, No. 5, pp. 477-500, Jan. 2001.
[6] J. Xu, T. Yamasaki, K. Aizawa, "3D Video Segmentation Using Point Distance Histogram," 電子情報通信学会 総合大会, D-12-14 Mar. 21-24 2005, Osaka.
[7] J. Xu, T. Yamasaki, and K. Aizawa, “3D Video Segmentation Using Point Distance Histograms,” Proceedings of 2005 IEEE International Conference on Image Processing (ICIP2005), Sep 11-14 2005 (accepted).
[8] D. Wang and C. Cui, “3D Model Similarity Measurement with Geometric Feature Map based on Phase-encoded Range Image”, in Advances in Multimedia Information Processing - PCM2004 (5th Pacific Rim Conference on Multimedia), LNCS333, pp.103-110, Tokyo, 2004.
[9] I. T. Jolliffe, Principal Component Analysis, ISBN: 0387954422, Springer-Verlag.
[10] www.hal.k.u-tokyo.ac.jp/~yamasaki/fit2005/nihon-buyou.gif.
[11] www.hal.k.u-tokyo.ac.jp/~yamasaki/fit2005/batting.gif.
[12] www.hal.k.u-tokyo.ac.jp/~yamasaki/fit2005/pitching.gif.