LH-010

3 次元表面位置合わせによる Time-Varying Mesh の動き解析

Motion Analysis of Time-Varying Mesh using 3D Spherical Registration

山崎 俊彦　　相澤 清晴
Toshihiko Yamasaki　Kiyoharu Aizawa
東京大学大学院情報理工学系研究科電子情報学専攻

1．まえがき

多視点カメラを用いた物体や人物の 3 次元形状復元は、コンピュータ・ビジョンの分野で永年にわたり研究されている[1]-[5]。実世界の静止 3 次元物体を多視点シルエット画像から復元するアイデアは、既に 1974 年に Baumgart によって提案されている[6]。その後、Kanade らによって静止物体のみならず動く 3 次元物体をモデリングするための専用スタジオが開発され[1]、動的 3 次元映像の生成が注目を浴びるようになった。その後、実時間 3 次元映像撮影システム[2]や弾性メッシュ法[3]・ステレオマッチング法[4][5]などによるモデルの高品質化・高精細化の検討がなされ、コンテンツとして閲覧に耐えうる品質の映像生成が可能となってきている。現在ではデータ容量、映像品質の観点から 3 次元モデルはポリゴンメッシュで表現されることが多い。本論文では、3 次元ポリゴンメッシュのシーケンスとして表現された動的 3 次元映像を、Time-Varying Mesh (TVM)と呼ぶことにする。

従来の 3 次元 CG アニメーションと比較した場合の TVM の最も大きな特徴の 1 つは、ポリゴンメッシュの頂点数や結線関係がフレーム毎に異なる点である。すなわち、各フレームは隣接フレームを考慮せず、全く独立に生成される。これは、被写体の多くが衣服を着た人体などの非剛体であり、頂点数や結線関係を保存したまま次フレームを生成することが極めて困難なためである。

現段階では TVM の生成自体が発展途上な技術であるため、取得・生成以外で報告されいている研究例は圧縮に関するものが殆どである[7]-[9]。しかし、今後大規模な TVM の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、検索・編集・インデクシングなどの映像管理技術が必要不可欠である。以上の観点から、筆者らはこれまで動きセグメンテーション[10][11]、キーフレーム抽出[12]、動き検索[13][14]、編集[15]などの要素技術を開発してきた。また、他の研究グループからの応用技術も若干数ではあるが報告されている[16][17]。

我々の従来研究[10]-[15]では、各フレームの 3 次元モデルから何らかの形状特徴ベクトルを生成し、検索や編集などの処理を行ってきた。このアプローチは処理コストが低いというメリットがある反面、逆に物体の 3 次元空間上での動きの情報は失われてしまう。文献[14]で議論されているように、今後 TVM に対してより正確な動き理解・処理を行っていくためには 3 次元空間上での動き解析が必要不可欠である。

本論文では、Iterative Closest Point (ICP)法[18]を用いた、メッシュ頂点の 3 次元表面位置合わせによって動きを解析する手法を開発した。この手法では、フレーム間の動き量、即ちフレーム間の非類似度を ICP 法のマッチング誤差によって表現している。この方式によって動きセグメンテーションと類似動作検索を実現した。特に動きセグメンテーションに関しては適合率 95%、再現率 92%と従来手法に比べて格段によい性能を発揮することが実験により示された。また、類似動作検索に関しても少ない実験データではあるが実現可能性を示すことができた。

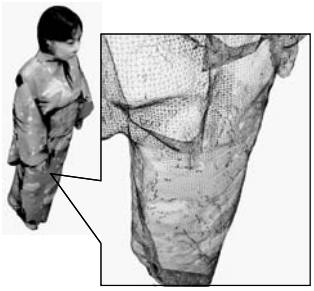

Fig. 1. TVM データの一例。

2．Time-Varying Mesh (TVM)

本論文で使用する TVM データは冨山らによって開発されたシステム[4]を用いて生成された。直径 8m、高さ 2.5m の専用ブルーバック・スタジオで、22 台の同期カメラを用いてモデル生成を行った。

TVM の各フレームは 3 次元メッシュモデルから構成されている。我々が扱う TVM の、あるフレームを 1 つの視点から見た例を Fig. 1 に示す。Fig. 1 から分かるとおり、各フレームは多数の頂点、頂点同士の結線情報、各頂点(または三角パッチ)の色の 3 種類の情報からなっている。言うに及ばないが、視点位置はユーザの好みによって任意に変えることができる。

3. アルゴリズム

3.1　フレーム同士の非類似度評価

従来の TVM 処理においては、フレーム同士の類似度は何らかの形状特徴ベクトル空間で計算されていた[10]-[15]。この方法は演算コストを低く抑えることが出来るが、動き解析の性能に限界があり、3 次元空間上での動き解析が必要であると指摘されてきた[14]。本論文では、ICP 法[18]を用いて 3 次元空間上での 3 次元モデルの動きを表現する手法を開発した。ICP 法はレーザスキャン・データなど複数の 3 次元点群(point clouds)同士の空間的位置あわせを行うのに広く用いられている手法である。本論文では、3 次元メッシュモデルの頂点データに対して ICP 法を用いて位置合わせを施し、その結果得られる位置合わせ誤差、

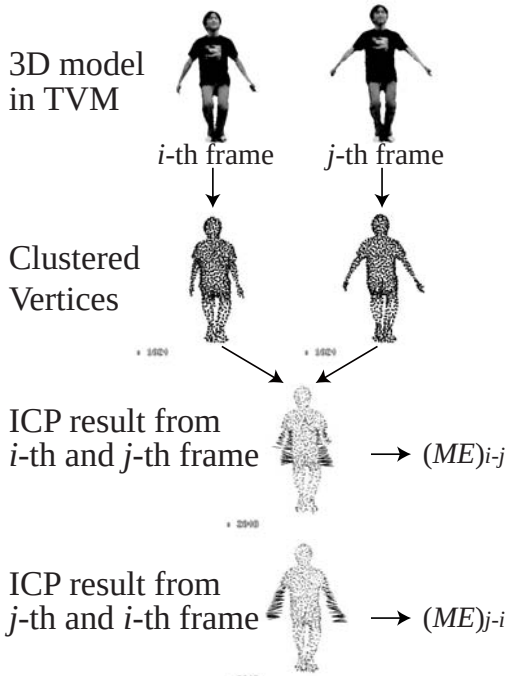

Fig. 2. ICP 法を用いたフレーム間の非類似度演算方法。

すなわち対応すると思われる頂点間の距離の総和をフレーム間の非類似度と定義する。対応頂点間の距離が大きいほど、大きく動きのある箇所であることを示している。

我々のデータベースに登録されている TVM データは、モデルの空間解像度(3.3mm～10mm)に依存するものの、各フレームにおおよそ 2 万～10 万点の頂点を含んでいる。ICP 法は頂点数に対して二乗のオーダーで演算コストが増大するため、数万点の頂点に対して処理を施すのは現実的ではない。そこで、本論文ではベクトル量子化によって頂点群を空間的に 1024 の領域にクラスタリングする。そして、各クラスタの重心ベクトルを用いて ICP 法を施すことにより演算コストの低減を実現する。頂点数を適切に削減して演算コストを削減するアイデアは[20]と類似しているが、頂点数削減の手法が異なる。

また、ICP 法による位置あわせは非対称である。すなわち、*i* 番目のフレームを *j* 番目のフレームに合わた場合と、その逆では頂点の対応関係が異なり、その結果計算されるフレーム間の非類似度も異なる。そこで、本論文ではその影響をなくすために *i* 番目のフレームと *j* 番目のフレームの非類似度(D_{i-j})を以下のように定義する。

$$D_{i-j} = D_{j-i} = (ME)_{i-j} + (ME)_{j-i} \quad (1)$$

ここで、$(ME)_{i-j}$ は *i* 番目のフレームを *j* 番目のフレームに合わせた場合の位置合わせ誤差を表現している。以上の処理の概略図を Fig. 2 に示す

3.2 動きセグメンテーションと類似動作検索

動きセグメンテーションは、長い動きシーケンスを動きの意味の区切れ毎に細分化し、その後に続く検索や編集などの処理をしやすくするための重要な前処理である。また、細分化された TVM シーケンスのことをクリップと呼ぶことにする。

Table 1. Parameters of 3D video utilized in experiments.

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

本論文では、セグメンテーションの候補位置を隣接フレーム間の動き量が極小となる時点を探索することで抽出する。*n* 番目のフレームの動き量は *n* 番目のフレームと *n*+1 番目のフレームの非類似度$(DM)_n$によって表現する。

$$(DM)_n = D_{n-(n+1)} \quad (2)$$

運動力学的特徴量が極小になる時点をセグメンテーション位置とする手法は、2 次元映像やモーションキャプチャ・データなど様々な動きデータに対して一般的に用いられている[11][21][22]。特に、舞踊やダンスなどの動きは、動きをより優雅に見せるために動きの意味の区切れで一時的に動作をとめたり動きを意図的に小さくすることが多いため、このアプローチは特に有効である。

しかしながら、単に動き量が極小となる時点を抽出する手法は検出漏れが少ない一方、過検出が多く発生するため、verification が欠かせない。従来手法[11][21][22]では、経験的に決定された閾値による閾値処理で verification を行っていた。しかし、動きの大きさはシーケンスによって大きく異なる。例えば、ヒップホップやブレークダンスなどのアクロバティックな激しい踊りは動きが大きい。一方、能などは優雅でゆっくりとした動きがおおい。そのため、動きの種類によらず一定の閾値処理をすることは非常に困難である。

そこで、本論文では筆者らが文献[11]で開発した、相対比較による verification 手法を導入する。この手法では、各極小点の直前と直後に起こる極大値に注目する。その 2 つの極大値がいずれも極小値のα倍よりも大きい場合に正しいセグメンテーション位置と見なす。また、この条件を満たさない極小位置は過検出(ノイズ)とみなす。なお、本論文ではαの値は 1.1 である。以上の処理を式で表現すると以下の通りである。

$$\begin{cases} \text{if } (lmax)_{\text{before}} > \alpha \times lmin \text{ and } (lmax)_{\text{after}} > \alpha \times lmin, \\ \quad \text{the local minimum point is a segmentation point;} \\ \text{otherwise,} \\ \quad \text{the local minimum is regarded as noise.} \end{cases} \quad (3)$$

ここで、*lmin* は極小値、$(lmax)_{\text{before}}$, $(lmax)_{\text{after}}$ はそれぞれ極小値の前後に発生する極大値を示す。

動きセグメンテーションによって細分化されたクリップは、次に続く類似動作検索において基本最小単位として扱う。本論文では、例示方式による類似動作検索を行う。すなわち、ある TVM クリップをクエリとして使用し、データベースに含まれる他のクリップから類似動作を検索する。各 TVM クリップに含まれるフレーム数は一致しないことが多い。そこで、筆者らが文献[14]で用いたように、動的計画法(DP マッチング法)によってクリップ同士の類似度を評価する。なお、フレーム間の非類似度は式(1)で定義した式を使用する。動的計画法によるクリップ間の類似度評価手法の詳細については文献[14]を参照されたい。

4. 実験結果

本論文の実験に於いては、文献[4]のシステムを使用して生成した、5 つの TVM シーケンスを用いた。データの諸元を Table 1 に示す。シーケンス#1 と#2-1～#2-3 は盆踊り、#3 はラジオ体操の動きである。シーケンス#2-1～#2-3 は同一の動きであるが、踊り手が異なる。動きセグメンテーションの正解は 8 人の被験者の主観的判断に統計的処理を施して生成した[11][14]。α値は特に断りのない限り 1.1 とした。

動きセグメンテーションの例として、シーケンス#2-1 のはじめの 21 秒間に対する実験結果を Fig. 3 に示す。動きの意味の区切れで正しくセグメンテーション出来ていることがわかる。また、図中×印があるものは過検出で、検出漏れはなかった。

Table 2 に動きセグメンテーションの性能をまとめる。文献[11]の、筆者らの先行研究による結果も比較のために掲載してある。先行研究に比べ recall rate (適合率)が高く、検出漏れが大きく改善されている様子が見て取れる。検出漏れの少なさは本手法において特に重要である。なぜなら、過検出は verification によってその多くを取り除けるのに対し、検出漏れは対処する手段がなく、最初に取りこぼし無く候補を検出する以外に方法が無いためである。全てのシーケンスに対する precision rate (適合率)、recall rate (再現率)、F 値はそれぞれ 95%、92%、94%であった。なお、F 値は適合率と再現率の調和平均であり、総合的な性能評価に用いられる手法である。

本論文の手法と従来手法によるセグメンテーションの性能比較を Fig. 4 に示す。適合率-再現率カーブは、それぞれの手法に於いて verification のための閾値パラメータを変化させることで得た。例えば、本論文の手法ではα値を 1.05 から 1.4 まで変化させた。

類似動作検索においては、シーケンス#2-1 と#2-2 それぞれから 5 種類の動きをれ 2 クリップずつ、合計 10 クリップ()を任意に選択した。選択したクリップは「手とたたく」、「両腕で大きな円を描く」、「体を右にねじる」、「３回ジャンプ」、「ジャンプして体全体で大の字を描く」である。例として、「両腕で大きな円を描く」の動きシーケンスを Fig. 5 に示す。

クリップ間の類似度を図示したものを Fig. 6 に示す。色が黒いほどクリップ間が似ていることを表している。似た動き同士は高い類似度を示していることがみてとれる。現段階ではまだ少ない数のクリップしか用いていないものの、類似動作検索の可能性を示すことができた。今後は大規模なデータベースに対して性能評価実験を行う予定である。

ICP 法による点群位置あわせは、非常に演算コストがかかるため実時間処理には向かないという問題がある。例えば、現在 1 フレーム当たりの処理時間は 1～3 秒程度かかっている。これは今後解決すべき問題である。

5. まとめ

本論文では、TVM の動きを 3 次元空間上で解析する手法を開発した。本手法では、フレーム間の非類似度を ICP 法による位置合わせ誤差の総和によって表現した。これにより、従来手法と比較してより高精度な動きセグメンテーションと類似動作検索が実現できた。

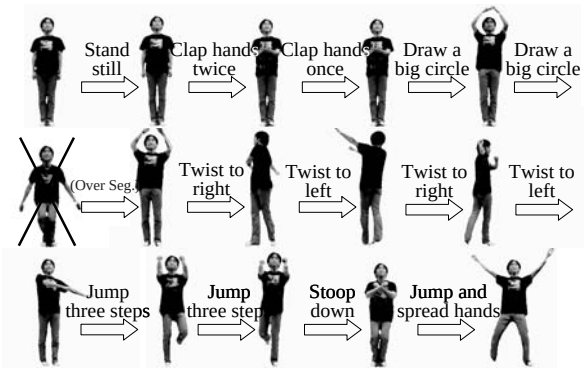

Fig. 3. シーケンス #2-1 に対する動きセグメンテーション結果。

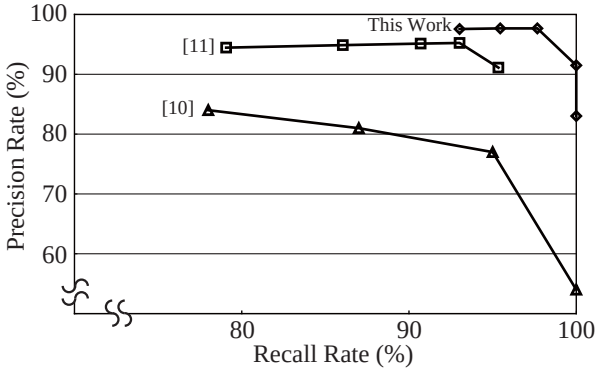

Fig. 4. シーケンス#2-1 に対する適合率-再現率曲線。

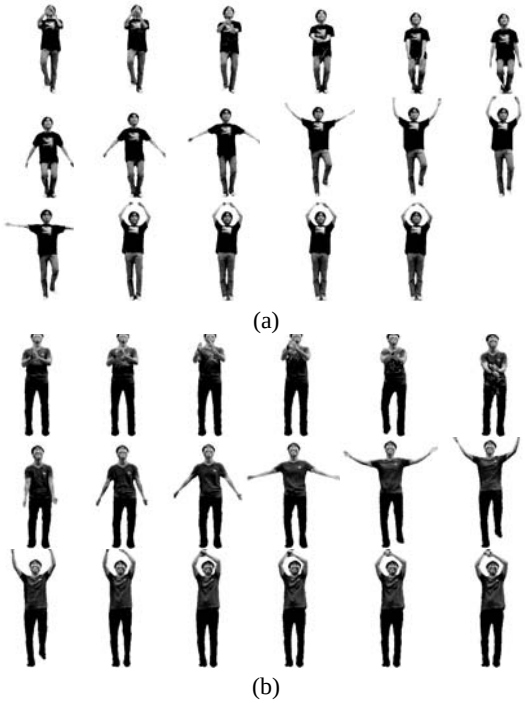

Fig. 5. 「両腕で大きな円を描く」の例: (a) #2-1, (b) #2-2.

Table 2. 動きセグメンテーションの性能まとめ

Sequence	# 1	# 2-1	# 2-2	# 2-3	# 3	Total	Total [11]
A: # of relevant records retrieved	10	44	46	41	127	268	251
B: # of irrelevant records retrieved	2	3	3	2	5	15	23
C: # of relevant records not retrieved	1	0	0	1	20	22	39
Precision (P): A/(A+B)	83.3	93.6	93.9	95.3	96.2	94.7	91.6
Recall (R): A/(A+C)	90.9	100	100	97.6	86.4	92.4	86.6
F value: 2PR/(P+R)	87.0	96.7	96.8	96.5	91.0	93.5	89.0

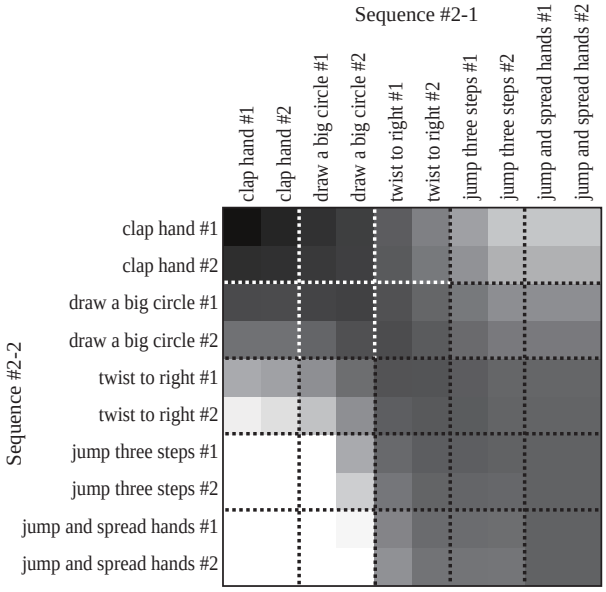

Fig. 6. クリップ同士の類似度評価結果。色が濃いほど類似度が高いことを示す。

参考文献

[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, Jan./March 1997.

[2] W. Matusik, C. Buehler, R. Raskar, S. Gortler, and L. McMillan, “Image based visual hulls”, ACM SIGGRAPH2000, pp.369-374, 2000.

[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 TCSVT, vol. 14, no. 3, pp. 357–369, March 2004.

[4] K. Tomiyama, Y. Orihara, M. Katayama, and Y. Iwadate, “Algorithm for dynamic 3D object generation from multi–viewpoint images,” Proc. SPIE, Vol. 5599, pp. 153–161, 2004.

[5] J. Starck and A. Hilton, “Virtual view synthesis of people from multiple view video sequences,” Graphical Models, vol. 67, no. 6, pp. 600-620, 2005.

[6] B.G. Baumgart, Geometric modeling for computer vision, Ph.D. thesis, Stanford University, 1974.

[7] H. Habe, Y. Katsura, and T. Matsuyama, “Skin-off: representation and compression scheme for 3D video,” Proc. Picture Coding Symposium (PCS2004), pp. 301-306, 2004.

[8] K. Muller, A. Smolic, M. Kautzner, P. Eisert, and T. Wiegand, “Predictive compression of dynamic 3D meshes,” Proc. IEEE ICIP2005, pp. I-621-I-624, 2005.

[9] S. Han, T. Yamasaki, and K. Aizawa, “3D video compression based on extended block matching algorithm,”, Proc. IEEE ICIP2006, pp. 525-528, 2006.

[10] J. Xu, T. Yamasaki, and K. Aizawa, “3D video segmentation using point distance histograms,” Proc. IEEE ICIP2005, pp. I-701-I-704, Sep 11-14, Italy, 2005.

[11] T. Yamasaki and K. Aizawa, “Temporal 3D video segmentation using modified shape distribution,” Proc. IEEE ICME2006, pp. 1909-1912, 2006.

[12] J. Xu, T. Yamasaki, and K. Aizawa, “Key frame extraction in 3D video by rate-distortion optimization,” Proc. IEEE ICME2006, pp. 1-4, 2006.

[13] T. Yamasaki and K. Aizawa, “Similar motion retrieval of dynamic 3D mesh based on modified shape distribution,” Proc. Eurographics2006 short papers, pp. 9-12, 2006.

[14] T. Yamasaki and K. Aizawa, “Motion segmentation and retrieval for 3D video based on modified shape distribution,” EURASIP Journal on Applied Signal Processing, vol. 2007, Article ID 59535, 11 pages, 2007.

[15] J. Xu, T. Yamasaki, and K. Aizawa, “Motion editing in 3D video database,” Proc. 3rd Int. Symp. 3D Data Processing, Visualization and Transmission (3DPVT), #96, 2006.

[16] J. Starck and A. Hilton, , “Spherical matching for temporal correspondence of non-rigid surfaces,” IEEE ICCV2005, pp. 1387-1394, 2005.

[17] Starck, J. and Miller, G. and Hilton, A., “Interactive free-viewpoint video,” European Conference on Visual Media Production, pp. 52-61, 2005.

[18] P.J. Besl and N.D. McKay: ”A method for registration of 3-D shapes,” IEEE TPAMI, vol. 14, no. 2, pp. 239–256, Feb 1992.

[19] T. Yamasaki and K. Aizawa, “Motion segmentation for 3D video based on spherical registration,” 3DTV-Conference, 2007. (submitted)

[20] Turk, G. and Levoy, M. “Zippered Polygon Meshes from Range Images,” Proc. SIGGRAPH94, pp. 311-318, 1994.

[21] T.S. Wang, H.Y. Shum, Y.Q. Xu, and N.N. Zheng, “Un-supervised analysis of human gestures,” Proc. IEEE Pacific Rim Conference on Multimedia, pp. 174-181, 2001.

[22] T. Shiratori, A. Nakazawa, and K. Ikeuchi, “Rhythmic motion analysis using motion capture and musical information,” Proc. IEEE MFI2003, pp. 89-92, 2003.