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

2008－SLDM－136（1）
2008 / 10 / 6

[招待講演] 3次元映像の圧縮と処理

相澤 清晴†

†東京大学 情報理工学系研究科 電子情報学専攻

〒113-8656 東京都文京区本郷 7-3-1

E-mail: †aizawa@hal.t.u-tokyo.ac.jp

あらまし 変化する実世界の人物などの対象をありのままに取得し，3次元映像として再生する試みが進んでいる．3次元映像とすることで，視点位置が自由になるばかりでなく，インタラクティブな操作，編集も可能になる．取得できる3次元映像の品質も向上し，今後の身近での利用が期待される．ただし，その広い利活用の展開のためには，圧縮をはじめとするさまざまな処理が必須である．本稿では，3次元映像の現況を概観するとともに，圧縮をはじめとする処理についてのわれわれの検討を紹介する．

キーワード 3次元映像，Time-Varying Mesh，圧縮，検索

Compression and Processing of 3D Video

Kiyoharu AIZAWA†

† Dept. of Information and Communication Engineering, University of Tokyo

7-3-1 Bunkyo Tokyo 113-8656 Japan

E-mail: †aizawa@hal.t.u-tokyo.ac.jp

Abstract We are working on processing of 3D video, which is a sequence of 3D models. 3D video reproduces a real moving object and provides free view point functionality. Differing to CG animation, models in the sequence of 3D video varies in the number of their vertices, connectivities, etc. The quality of capture and reproduction of 3D video is now very much improved. We further work on new topics such as compression, retrieval, editting, etc of 3D video for its wider applications. In this paper, quickly summarizing the status of the art of 3D video and we would like to present our work focusing on compression, retrieval etc.

Keyword 3D video, Time Varying Meshes, Compression, Retrieval

1. はじめに

3次元映像技術の研究開発は，長く盛衰を繰り返してきたが，映像のデジタル技術の進んだ現在，世界的に歩調を合わせて盛んになってきている．欧州では，3DTVという19機関の参加する大きなプロジェクトが2004年から始まり，2008年から3D Media Clusterへと進展している．米国では，3D@Homeなるコンソーシアムが2008年に起こり産業化を指向し，日本では超臨場感コミュニケーションフォーラムが2007年に始動した．これらの取り組みは，次世代のメディアとしての3次元映像への期待を強くあらわしているものと思われる．

今，身近なテレビで目にする映像は，そのカメラやディスプレイのもとをたどれば，1世紀ほど遡る．その研究開発の歴史の中で，入出力のカメラ，ディスプレイのみならず，放送，通信，数えきれないアプリケーション，そのための圧縮や処理技術が生まれてきた．後述するように3次元映像の取得は90年代の半ばにその試みが始まったと考えていい．すると，3次元映像の現在は，ふつうの2次元映像で例えれば，そのカメラができて間もないころに相当する．我々は，これから，3次元映像技術が離陸するのか，しないのかのまさに節目あたりにいるのではないだろうか．取得の品質は向上している．3次元映像が，今後，広く利活用されるためには，圧縮をはじめとした処理が重要であると考えている．

本稿では，時間的に変化する幾何データに基づく3次元映像に関してのわれわれの取り組みについて述べたい．

2. 3次元映像

そもそも3次元映像とは何なのか．コンセンサスのとれた定義はないように思われる．ディスプレイの方式にとらわれずに，整理すれば，以下の要請を満たすようなものであろうと考えている．

－実写であること
－動的であること
－自由視点であること

最後の点が，通常の2次元映像に対して異なるところであり，最初の点が純粋なＣＧとの差になる．

このような条件を満たす3次元映像の表現に関しては，大きくわけて2つアプローチが知られている．それらは，

－イメージベース　（例えば，[2,3]）
－幾何モデルベース　（例えば，[4-6]）

のアプローチである．興味深いことに，いずれの手法も実際の動く対象に対しての検討が進んだのは，90年代の半ばの時期である．いずれの手法ともその取得はまだ容易とはいえない．イメージベースのアプローチでは，眼数が極めて多くカメラを必要とする．モデルベースのアプローチでは，通常，複数視点からのカメラ画像でモデル生成し，そのモデル生成が容易ではない．ただし，後者に関しては，いったんモデルができれば，そのあとの編集加工にあたっての自由度は極めて大きく，再利用が容易になる．

我々は，後者のアプローチを選ぶ．そのモデルは，視体積交差法とステレオマッチングを併用する手法により生成されたものである．各フレームのモデルは独立に作られ，3次元のモデル時系列として，3次元映像を表現することになる．モデルは，富山らの手法[4]で生成されており，品質の高い表現になっている．そのモデルの頂点数は，大よそ，約50,000～100,000頂点であり，ファイルサイズとしては，一フレームあたり，5MB～10MBもの大きさがある．（仮に1分間のシーケンスとした場合，3-5GBにもおよぶ．このため，圧縮は必須である．）

なお，ごく最近，モデル生成に関して，テンプレートを変形するという視点での研究がSIGGRAPH2008にて2件[7,8]発表され，精度を向上する試みが進んでいる．

3. 3次元映像の圧縮

3次元幾何データの圧縮は以下のようなカテゴリに分けられる．

(1) 静的3次元メッシュの圧縮
　時間的に変化しない3次元オブジェクトの圧縮．
(2) 動的3次元メッシュの圧縮
　接続関係などは時不変のまま，頂点位置が変化する．コンピュータアニメーションは通常これに該当する．
(3) 3次元メッシュ時系列の圧縮
　動的な3次元メッシュであるが，各フレームのモデル間での頂点対応はなく，頂点数も接続関係も時間的に変化する．

これまで幾何データの圧縮に対する取り組みは多いものの，そのほとんどが（1）（2）のケースを扱ってきた．静止3次元メッシュと動的3次元メッシュの圧縮に関しては，MPEG4における3D Mesh Compressionでの標準化もなされている[1]．

一方，(3)は，もっとも自由度の高い表現となっている．我々の興味の対象となっている3次元映像は，フレームが独立に生成されることから，Mesh自身が頂点数まで含めて時間的に変化するため，(3)に該当し，圧縮にとって新しい課題であると考えている．3次元映像の生成自体が比較的新しい課題であり，対象となるデータも多くないことから，その圧縮はほとんど扱われてこなかった．なお，以降，この(3)に相当する3次元データをTime Varying Mesh(TVM)とも称することにする．

効率的な3次元映像の圧縮を行うためには，2次元の動画像圧縮と同様にフレーム内圧縮とフレーム間圧縮の組み合わせが有効である．
我々はこれまで，3次元映像を効率よく圧縮するために，大きく二つのやり方を検討してきた．その一方は，2次元映像で有効なブロックマッチング手法を3次元映像に適用した拡張ブロックマッチング[10]であり，もう一方は，2段階量子化とランレングス符号化による手法[11]である．前者は，計算量が多い割に，オバーヘッドが大きい．後者は，比較的簡単な処理ながら，圧縮効率に優れている．以下，後者に関して概略を述べる．

◆2段階量子化とランレングスによる圧縮[11]

提案手法の説明のため，N個の3次元モデルで表現されるTVMの時刻tでの3次元モデルを$M(t)$, t=1,2, …, Nとする．まず前処理として，各3次元モデル$M(t)$の質量重心点が原点にな

るように，$M(t)$の平行移動を行う．次に，平行移動された $M(t)$を囲む共通バウンディングボックス B を求める．B の x, y, z 軸の一辺の辺長さ l_x, l_y, l_z を長さ s で切り，$M\times N\times L$ 個のキュービックブロック群 $C^p(t)$を求める．$C^p(t)$には $M(t)$の部分表面を含むブロック $S^k(t)$と含んでいないブロック $R^q(t)$が存在する．図１にブロック化結果を示す．$C^p(t)$は $M(t)$の低解像度表現であるため，連続するモデルの形状変化を容易に検出することができる．

$C^p(t)$の圧縮のため，$S^k(t)$にシンボル"1"を，$R^q(t)$にシンボル"0"を割り当て，バウンディングボックス B 内を x, y, z 軸順に走査しながら，$C^p(t)$のバイナリーストリーム化を行う．TVM の空間的冗長度により，$C^p(t)$中には $R^q(t)$の割合が多いため，$C^p(t)$のバイナリーストリーム $B_{bin}(t)$は長い"0"のランが発生する．そのため，$B_{bin}(t)$はランレングス符号化で効率よく圧縮できる．

TVM の時間的冗長度を用いて，$C^p(t)$の更なる圧縮を図る．連続するモデル $M(t\text{-}1)$, $M(t)$を考える．それぞれのブロック化結果を $C^p(t\text{-}1)$, $C^p(t)$，バイナリーストリームを $B_{bin}(t\text{-}1)$, $B_{bin}(t)$とする．時間的に隣り合う $C^p(t\text{-}1)$, $C^p(t)$は類似していることから，$B_{bin}(t\text{-}1)$, $B_{bin}(t)$の排他的論理和を取ることで，モデル間形状の変化がある場所のみが"1"が発生し，他のところはすべて"0"になる．その結果，$B_{bin}(t)$はさらに長い"0"ランが発生され，更なる $C^p(t)$の圧縮ができる．

次に，$S^k(t)$に含まれている頂点情報 $V^k(t)$の圧縮を行う．まず，$S^k(t)$の空間を格子状に量子化を行う．量子化誤差は量子化ステップ Q を十分細かくすることで無視できる．その後，$V^k(t)$の量子化位置を表す量子化標本点にシンボル"1"を，それ以外の標本点にシンボル"0"を与え，x, y, z の順に $S^k(t)$空間内を走査しながら，頂点情報のバイナリーストリーム化を行う．TVM の空間的冗長度の性質から，頂点情報のバイナリーストリームは長い"0"ラン発生する．$V_{bin}{}^k(t)$はランレングス符号化を用いて効率よく圧縮できる．

我々は，173 フレームからなる TVM シーケンスを用いて圧縮実験を行った．実験では，$C^p(t)$の一辺の長さは $s=\max(l_x, l_y, l_z)/2^4$ に設定した．$C^p(t)$内の量子化は一様量子化で，量子化ステップは，$Q=s/2^5$ に設定した．図２に圧縮率と誤差を示す．誤差は圧縮前後の頂点位置の距離の差の平均を用いた．圧縮前の頂点座標は 96 bpv が必要であることに対して，提案手法により，5.6-6.1 bpv まで圧縮することができた．また，誤差は 0.15-0.17cm であり，ダイナミックレンジ 200cm 前後である TVM シーケンスに対して，誤差はほぼ無視できる．また，ビットレートを変化させたときのデコード結果を図１に示す．

図１．３次元モデルのブロック化．

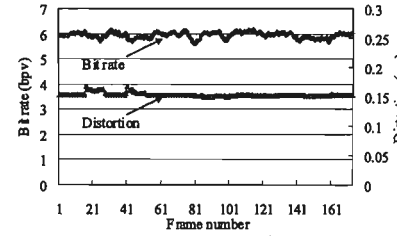

図２．提案手法によるビットレートと誤差．

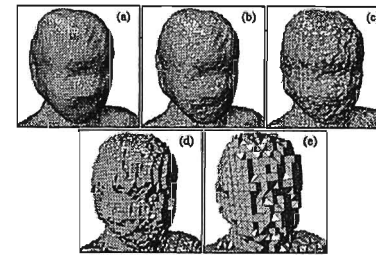

図３．顔の部分のデコード結果のクロースアップ．

4. 3 次元映像の処理

３次元映像は，今後，大規模に利用できるようになるとすれば，それを扱う処理技術が必要になる．そのために，以下にあげる処理についての検討を進めてきた．

－特徴抽出 [12,13,18]等
－時系列のセグメンテーション(分節化）[12,13]等
－キーフレーム抽出　[14]等
－３次元映像間の類似検索[13]等
－モーションキャプチャデータからの類似動

作検索 [15]等
－スケルトン抽出と人体モデルのセグメンテーション [16]等
－スケルトンの動きデータに基づくモデルベース符号化 [17]等
－ 3 次元映像の編集[18]等
以下に，そのいくつかを述べる．

4.1 セグメンテーションと類似検索[13]

◆セグメンテーション

我々は 3 次元映像のセグメンテーションについて取り組み，これまでに

- 複数基準点から各頂点までの距離ヒストグラム
- 極座標表現による距離・角度ヒストグラム
- バウンディングボックスの体積変化解析
- Modified Shape Distribution 法による形状・姿勢特徴表現

などの手法を提案してきた．いずれも，画像に対するヒストグラムのように，グローバルな特徴量を用いて，時系列のセグメンテーションに利用している．

例えば，Shape Distribution 法という形状に対する特徴量を修正し，Modified Shape Distribution 法を提案し，その時間変化を用いたセグメンテーションを行った．Modified Shape Distribution 法では，頂点のクラスタリングを行うことで，もともとの手法に比べて，格段に安定に形状特徴を求めることができる．

まず 1 フレームずつの 3 次元モデルの形状・姿勢を Modified Shape Distribution により，特徴ベクトル化する．その特徴ベクトルを用いて隣接フレーム間の距離を計算し，変化が極小値になった時点をセグメンテーションの位置とした．その結果，適合率 0.84，再現率 0.80 という他手法と比べて遜色のない結果が得られた．形状変化の極小値は，一般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ことが多く，いわば，動きの意味の句切れとなる「間(ま)」に注目した手法である．

Modified Shape Distribution 法を用いた時のセグメンテーション結果を図 4 に示す．この場合，検出漏れはなく，過検出が 3 件生じた．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も言えるが，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特徴量の信号処理的アプローチを取っているので必ずしも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．

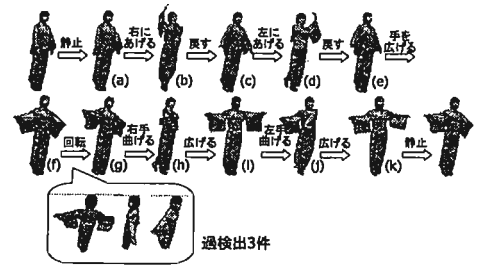

図 4. 3 次元映像のセグメンテーション(シーン分割)結果．図中(a)～(k)は 8 人の被験者による主観的セグメンテーション結果を示す．

◆類似動作検索

類似動作の検索においては，クエリとなるシーケンスと検索対象となるシーケンスのフレーム数が等しいとは限らない．そのような場合，Dynamic Programming (DP) マッチングによる検索が有効である．DP マッチングとは長さの異なる 2 つのシーケンスに対してどちらか一方をずらしながらシンボル間の一致・不一致またはずらし量に応じてコストを加算していき，累積コストが最も小さくなるような「ずらし」を探索する手法である．フレーム間の類似度評価は Modified Shape Distribution 法で生成したヒストグラム間のユークリッド距離を用いる．そして，最終的な累積コストを 2 シーケンス間の距離として定義する．ただし，累積コストはシーケンス長に依存するので 2 つのシーケンスのうち長い方のフレーム数でコストを正規化する．また，セグメンテーションによって分割されたシーケンスそれぞれを 1 つの単位として扱うことで効率的な検索を実現する．図 4 に示した各シーケンスに対する DP マッチングによる類似度評価の結果を図 5 に示す．

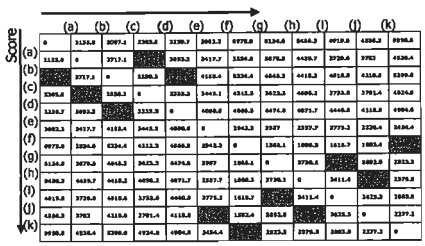

図 5. DP マッチングによる各シーケンスの類似度

類似動作検索の代表例を図 6 に示す．なお，図 6 に示した例は検索の結果もっとも類似しているシーケンスのみを掲載している．実験の結果，この 3 次元映像に対しては 100%人間の主観評価と一致する結果が得られた．

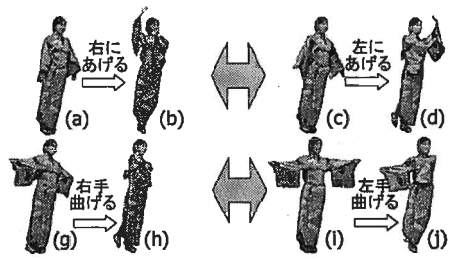

図 6. 類似動作検索結果の一例

4.2. セグメンテーションとスケルトン抽出[16]

3 次元映像は前述の通り各フレームが独立して生成されることが多いため，隣接フレーム間での頂点や結線情報同士の対応がとりにくく，そのため様々な処理が困難となっている．そこで，その問題を解決するための手法として高精度なモデル・セグメンテーションとスケルトン抽出技術を開発した．この手法では，衣服をきたままで，スケルトンの抽出とその動きの把握が可能である．

我々の手法は 3 つの要素からなる．3 次元映像からの人体構造モデル抽出，階層的モデル分割，スケルトン抽出・高精度化である．各フレームの初期スケルトンは，それよりも前の数フレームを用いたベジエ曲線による動き予測により生成する．階層的モデル分割は 3 次元メッシュモデルを予め定義された複数のパーツに正しく分割するための処理である．最初に，3 次元モデル上の各頂点は最も距離の近いスケルトンと対応付けされる．モデルは第一段階として 6 つのパーツ(頭，両手，両足，胴体)に分割され，第二段階でさらに細かいパーツへと分割される(図 7 参照)．その後，動き予測によって得られたスケルトンはセグメンテーション結果に対応するよう，変形・微調整される．セグメンテーションとスケルトン微調整を交互に繰り返すことにより，高精度化を行う．

また，セグメンテーションの精度をさらに向上させるため，セグメンテーションされた部分の色を考慮することもある．ただし，実験の結果色情報を用いたセグメンテーションの高精度化はヒップホップなど動きが激しく，隣接フレーム間の対応をとるのが特に困難である場合にのみ必要であり，それ以外の場合には特に必要がないことも明らかとなっている．

セグメンテーションの安定性は，セグメンテーションされた各パーツの表面積率の変化によって評価した．ヒップホップの激しい動きに対しては 120 フレームに対して平均変化率 2.6%，動きの小さなダンスに対しては最大変化率 0.92%と非常に安定した性能を実現することができた．

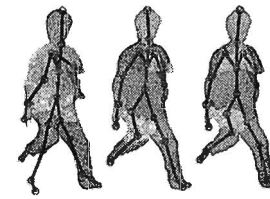

図 7．モデルのセグメンテーションとスケルトン抽出．

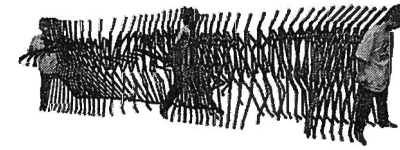

図 8．スケルトンを用いた動き追跡．

4.3. モデルベース符号化への利用[17]

シーケンスから取り出したスケルトン情報に従い，TVM 中のある 1 フレームを変形・制御する事で，圧倒的に少ないデータ量での 3 次元映像合成をおこなうことができる．その予備検討を以下に記す．

まず各フレームを胴体，頭，左右の上腕と前腕，左右の大腿と下腿の 10 個のパーツに分割し，そこからスケルトン情報を抜き出し，各フレームにおけるそれぞれの関節の回転情報を求める．そして，その回転情報を元に最初のフレームの各関節を回転させることにより，TVM に近い動きを表現させる．

図 9 に TVM とスケルトンを用いて最初のフ

レームを動かしたモデルをいくつか示す.上がTVM，下がモデルベース合成によるシーケンスである．各関節に若干違和感があるものの，映像としては視聴に耐えうるものになっている．シーケンスは 10fps の 25 フレーム分である．TVM は人間のモデルであれば最低でも空間解像度が 10mm 必要であり,データ量はおよそ 2MB/フレームとなる．よって TVM を普通に表示させると，25 フレーム，つまり 2.5 秒で 50MB 必要となる．一方，スケルトンデータは 1 フレームあたり 240 バイト程度なので 25 フレーム分でもたった 6KB であり，これと最初のフレームの 2MB だけで動きを表現できるので,データ量は大幅に削減できる事が分かる.

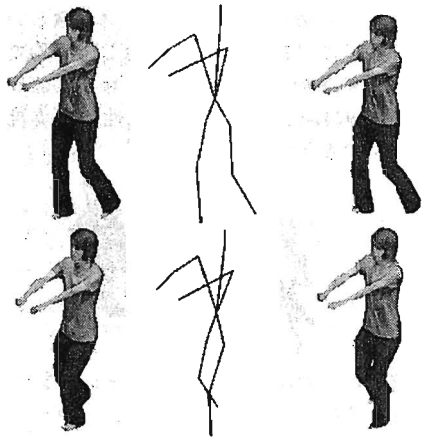

図 18. 動作合成結果．(左)オリジナルの 3 次元メッシュモデル，(中央)抽出したスケルトン，(右)第一フレーム目を該当フレームの姿勢に変形した結果.

5．まとめ

本稿では，新しいメディアとして期待される 3 次元映像に関して，その利活用を促進するための技術となる圧縮をはじめとする処理手法について，我々の取り組みを論じた．より簡易な取得環境やケータイ端末を含めたアプリケーションの広がりにより，一層身近なものになることを期待したい.

謝辞

3 次元映像の研究は，山崎俊彦講師をはじめとする研究室メンバと進めてきた.また,なお，文科省リーディングプロジェクトの中の課題「伝統舞踊の 3 次元映像アーカイブ」にてＮＨＫ岩舘氏，ＡＴＲ内海氏らと協力してきたことを付記し，謝意を表す.

参考文献

[1] Special Issue on Multiview Video Coding and 3DTV, IEEE Trans. CSVT, Vol.17, No.11, Nov. 2007.

[2] M. Levoy and P. Hanrahan, Light Field Rendering, SIGGRAPH '96

[3] T.Naemura, J.Tago, and H.Harashima: ''Real-Time Video-Based Modeling and Rendering of 3D Scenes," IEEE Computer Graphics and Applications, vol. 22, no. 2, pp. 66 - 73 (2002)

[4] T. Kanade, P. Rander, and P. Narayanan: Virtualized reality:Construction Virtual Worlds from Real Scenes, IEEE Multimedia, 4, 1, pp. 34-47, Jan./Mar. (1997)

[5] K. Tomiyama, Y. Orihara, M. Katayama, and Y. Iwadate, Algorithm for dynamic 3D object generation from multi-viewpoint image, Proceeding of SPIE, 5599, pp. 153-161 (2004)

[6] 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. CSVT, 14, 3, pp.357-369, Mar. (2004)

[7] J.Starck, A.Hilton, Surface Capture for Performance-Based Animation, IEEE Computer Graphics and Application, pp.21- 31, May/June 2007

[8] D.Vlasic, et al., Articulated Mesh Animation from Multi-view Silhouettes, SIGGRAPH2008

[9] E. de Aguiar, et al, Performance Capture from Sparse Multi-vew Video, SIGGRAPH2008

[10] S-R.Han, T.Yamasaki, K.Aizawa,Time-Varying Mesh Compression Using an Extended Block Matching Algorithm, IEEE Trans. CSVT,Vol.17, Issue 11, pp. 1506-1518, Nov. 2007

[11] S-R.Han, T.Yamasaki, K.Aizawa,Geometry Compression for Time-Varying Mesh Using Coarse and Fine Levels of Quantization and Run-Length Encoding, IEEE ICIP2008

[12] 徐，山崎，相澤，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,情報処理学会論文誌コンピュータビジョンとイメージメディア,Vol.47, No.SIG10 (CVIM15) pp.208-217, July, 2006

[13]T. Yamasaki and K. Aizawa, "Motion Segmentation and Retrieval for 3D Video Based on Modified Shape Distribution," EURASIP J. Applied Signal Processing, vol. 2007, Article ID 59535, 2007.

[14] J. Xu, T.Yamasaki, K.Aizawa, Summarization of 3D Video By Rate-Distortion Trade-off, IEICE Trans. Information and Systems,Volume E90-D, Number 9 Pp. 1430-1438, Sep. 2007

[15] T. Yamasaki and K. Aizawa, "Content-based cross search for human motion data using time-varying mesh and motion capture data," Proc. IEEE ICME2007, pp. 2006-2009, 2007.

[16] N.S. Lee, T.Yamasaki, K.Aizawa, Hierarchical Mesh Decomposition and Motion Tracking for Time-Varying-Meshes, IEEE ICME2008 ,pp.1565-1568

[17] T.Maeda, T.Yamasaki, K.Aizawa, Model-Based Analysis and Synthesis of Time Varying Mesh, Articulated Motion and Deformable Objects (AMDO) 2008, LNCS5098, pp.113-121

[18] J.Xu, T.Yamasaki and K. Aizawa, Motion Editing for Time-Varying Mesh, EURASIP Journal on Advances in Signal Processing,"Volume 2009, Article ID 592812, (2008.3)

[19] R. Tadano, T. Yamasaki, and K. Aizawa, Fast and robust motion tracking for time-varying mesh featuring Reeb-graph-based skeleton fitting and its application to motion retrieval, IEEE ICME2007, pp. 2010-2013, 2007.