研究速報

フルハイビジョン映像からの音楽演奏時の身体・手・顔のモーションキャプチャ

Capture of Body, Hands, and Face Motion from Full High-definition Television Images

渡部直人[†*], 糸雅亮祐[†], 佐藤真悟[†],
三浦裕樹[†], 正会員 山本正信[††]

Naoto Watanabe[†*], Ryousuke Itoga[†], Shingo Satou[†], Yuuki Miura[†] and Masanobu Yamamoto[††]

Abstract We propose a system that can capture the motion of an entire body from full high-definition (HD) television images. It is possible to precisely measure the movements of the body parts, i.e. the facial expression, trunk, arms, legs, and fingers using a high-resolution image. Processing high-resolution images requires a reduction of the computational efforts. We prepare a low-resolution image by downsizing an HD image to a quarter width and height, since capturing the motion of a person's trunk, arms, and legs does not always require a high-resolution image. After capturing the motion of the body from the low resolution image, we extract the images of the face and left and right hands from the HD image based on the captured motion so that each body part could be located in the center of the extracted images. The finger motion and facial expression are obtained from the HD quality images.

キーワード：モーションキャプチャ，ハイビジョンカメラ，顔・手・身体

1. まえがき

映画やＣＧアニメーション，バーチャルリアリティ，3Dゲームなどの分野では，モーションキャプチャの需要が高まっている．市販のモーションキャプチャの多くは，身体にマーカーなどを装着して動作を測定している．そのため，動作が制限される，コストが高いなどの欠点がある．これに対し，ビデオ映像のみを使用する画像処理方式では，身体に負担をかけずに自然な動作を手軽に測定することができる．

これまで，身体の動作，指の動き，顔の表情を個別に測定する画像処理方式は数多く提案されている．しかし，身体の動作測定に加えて，顔の表情や指の動きまで同時に測定することはなかった．それは，得られる画像の解像度が低かったためである．解像度の問題を解消するためにLokeら[1]は4台の同期したカメラを使って，顔の表情や指の動きを含む全身の動作測定を行っている．しかし，この方法では，手の素早い動きにカメラを追従させることが困難であり，また手軽に行えるモーションキャプチャシステムの構築を目指す上で，この環境を実生活で利用するのは現実的ではないと考えられる．

近年，ハイビジョンカメラの普及に見られるように，高解像度の映像が手軽に撮影できるようになってきた．そこで本研究では，1台のフルハイビジョンカメラからの映像を使って全身の動作を測定するシステムを提案する．

2. システムの構成

身体は縦長なので，ハイビジョンカメラを横に回転させ撮影する．画素数は，横1080画素，縦1920画素となる．なお，対象とする身体の奥行き方向の動きは小さいとして，カメラモデルは弱透視変換を仮定した．

ハイビジョン映像はサイズが大きいため処理時間の節約が課題となる．腕や脚の動きの測定では，VGA(640 × 480画素) サイズの映像で充分である．そこで，腕や脚の動きを測定するときは解像度を落とし，指や顔に対しては高い解像度で測定すれば，処理効率を上げることができる．

2011年6月20日受付，2011年7月25日再受付，2011年8月22日採録
†新潟大学 自然科学研究科
（〒950–2181 新潟市西区五十嵐2の町8050,TEL 025-262-7428）
††新潟大学 工学部
（〒950–2181 新潟市西区五十嵐2の町8050,TEL 025-262-7428）
*NTT コムウェア (株)
†Graduate School of Science and Technology, Niigata University
(8050 Ikarashi 2-no-chou, Nishi-ku, Niigata-shi 950–2181, Japan)
††Faculty of Engineering, Niigata University
(8050 Ikarashi 2-no-chou, Nishi-ku, Niigata-shi 950–2181, Japan)
*NTT COMWARE Co.

撮影したフルハイビジョン (HD) 映像から，測定対象となる部位に対応する四つの VGA サイズの測定用動画像（身体動画像・右手動画像・左手動画像・頭部動画像）を用意し，それぞれから動作を測定する．具体的には，まず HD 映像を縦横 1/4 に縮小することで身体動画像を生成し，身体の動作を測定する．**図 1** の中心の画像がハイビジョン映像で，縮小された VGA サイズの身体動画像が左上に示されている．このとき，解像度が低いため指の動きや顔の表情は測定しない．次に，測定された身体の動きを使って，対象部位が画像の中心になるように，手と頭部の動画像を HD 映像より切り出す．図 1 の左下，右下，右上にそれぞれ左右の手動画像と頭部動画像を示す．得られた手の動画像から手・指の動きを測定し，頭部動画像から顔の表情を測定する．最後に四つの動作データを統合し，全身のアニメーションを作成する．

画像サイズの解像度を制御することによる処理の効率化は既に試みられている．例えば，Bergen ら [3] による剛体の運動パラメータの推定，Wu ら [4] による時空間画像検索などがある．これらは画像間照合を解像度を変えて行うことにより，測定精度の向上や検索の効率化が図られている．本論文ではまず粗い画像から身体動作を測定している．さらに，詳細画像から測定精度を上げることも可能であるが，測定誤差が必ず含まれ得られた動作にドリフトを生じる．粗い画像から得られる動作の測定誤差を修正することにより，ドリフトの解消 [2] を行っている．一方，手や頭部など身体の末端部は，親部位の動きが加算されるため大きな動きを生じ易い．したがって，多重解像度による測定を行うべきところであるが，切り出された手の動画像，頭部動画像では手や頭部が画像の中心に保たれている．そのため，手に対する指や頭部に対する目や口の相対的な動きは小さく，多重解像度を用いずとも最大解像度での測定が可能である．

提案システムでは，身体と両手の動作及び顔の表情測定には既存の手法を使用している．本研究の新規性は，HD 映像から身体の階層構造を利用して指先から顔の表情に至るまで全身の動作を同時に測定することにある．

3. 身体と手の動作測定

画像式のモーションキャプチャでは，一般に，画像に身体の多関節モデルを照合させることで，身体の 3 次元姿勢を取得することができる．このとき，動画像の 1 フレームごとにモデルを照合させる必要はない．先頭フレームで身体モデルを照合し身体の姿勢を得た後，身体のフレーム間の動きはフレーム間差分画像から推定でき，推定した動作を先頭フレームで得た姿勢にフレーム順に累積して行けば動作を容易に測定することができる．本論文では，差分画像に基づく方法 [2] を使って身体および手の動作を測定する．

3.1 身体と手のモデル

身体の構造と手の構造を多関節モデルで表す．身体の多関節モデルを，**図 2** 左に示すように全部で 16 個の部位から構成し，腰部を根部位とする木構造で表す．各部位は根部位から頭や手，足に向かって親子関係で関係付けられる．

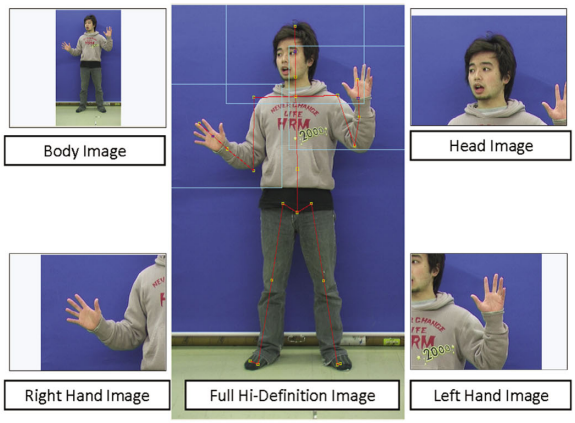

図 1 測定用画像作成　中央：ハイビジョン画像，左上：身体画像，右上：頭部画像，左下：右手画像，右下：左手画像，
VGA images for capturing motion of the body parts.

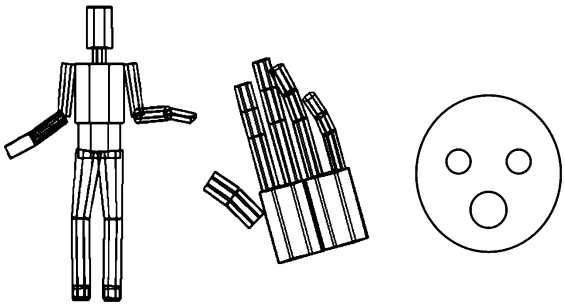

図 2 身体・手・顔のモデル
Left:Articulated model of human body, Center:Articulated model of hand, Right:Head and face model.

また，手のモデルは掌を掌 0 と掌 1 の二つに分け，親指，人差し指，中指は掌 1 の，薬指，小指及び掌 1 は掌 0 の子部位とする．掌 0 を木構造の根部位とする．手部位の総数も 16 部位である．図 2 中に示される左手のモデルは，掌が真ん中から二つに分かれ右が掌 0，左が掌 1 である．各部位の親子関係は掌 0 から指先に向かって関係付けられている．以下の記述では，掌は掌 0 を指すものとする．

なお，身体モデルと手のモデルは，それぞれ解像度の異なる画像で使用するため，身体モデルの手は一つの剛体で表され，手モデルのような詳細な指構造を含んでいない．

身体モデルも手のモデルも根部位の親はカメラ座標系とする．それぞれの部位は固有の座標系を持ち，部位の位置姿勢は直接の親部位に対する位置姿勢とする．位置は座標系原点の座標値，姿勢は x-y-z オイラー角で表す．部位 i の親部位 j に対する位置姿勢を同次変換行列 ${}^{j}T_{i}$ とする．この変換行列を使えば任意の部位から他部位への座標変換が容易に計算できる．例えば，身体の腰部，胸部，上腕，下腕，手を部位 $1, 2, \cdots, 5$ とすれば，手部位座標系からカメラ座標系への変換 ${}^{c}T_{5}$ は，

$$ {}^{c}T_{5} = {}^{c}T_{1}\,{}^{1}T_{2}\,{}^{2}T_{3}\,{}^{3}T_{4}\,{}^{4}T_{5} \qquad (1) $$

として計算される．ここで c はカメラ座標系を表す．

3.2 身体と手の動作測定

動作の測定では動画像から画像フレームごとに部位の位置と姿勢を計算する．先頭フレームにおいて，多関節モデルを身体像に照合させる．モデルと身体像の照合には様々な手法が提案されているが，いずれも身体像や背景，姿勢に制約がある．本論文では労力がかかるが手動で照合を行った．先頭フレームで得られた姿勢に，フレーム間での動きを逐次累積して行けば動作が自動的に得られるはずであるが，誤差も累積しドリフトを生じる．そこで，最終フレームや中間の幾つかのフレームでも姿勢を与え，累積姿勢が与えた姿勢と一致するようにフレーム間動きを修正しドリフトの解消を行う[2]．

4. 顔の表情測定

顔の構造を頭部を親部位，目と口を子部位とする．子の親部位上の位置は固定し，顔の向き及び目と口の開閉を測定する．図2右のように頭部を球体で表し，目と口を頭部の子部位とした．

4.1 顔および目，口の検出

頭部画像から顔や両目，口の位置検出は，OpenCVで公開されているHaar-Like特徴[5]を使用した．頭部部位（顔・両目・口）が充分に収まる大きさの部位画像を得る．矩形のサイズは先頭フレームでの検出サイズを基準とし，すべてのフレームで同じ大きさとする．具体的には，まず顔の位置を検出し，次に，顔画像内において，両目と口の位置を検出する．

検出ができなかったフレームも存在する．現在，未検出や誤検出を自動的に判定することはできず，検出結果を目で見て確認している．誤検出は手動で訂正し，未検出は，前後のフレームで検出された位置から線形補間により推定する．

目や口の部位画像に対して，各特徴点座標を求める．特徴点を，各部位の重心・上端・下端・右端・左端の五つとした．まず部位画像から肌色領域を検出し二値画像を得る．このとき，色空間をYCbCr表色系で表し，閾値処理により二値化を行った．ただし，閾値は撮影状況や個人差に依るので，画像ごとに変更を要することもある．

次いで得られた二値画像から面積が最大の連結領域を目や口の領域とする．それらの領域に対し，重心と二つの慣性主軸を計算する．二つの慣性主軸と領域の境界との交点をそれぞれ部位の右端・左端，上端・下端とする．上下の両端間の距離をあらかじめ決めておいた閾値と比較することにより開閉判定を行った．

4.2 顔の向きの測定

頭部を球体モデルで近似する．頭部動画像の中から顔が正面を向いているフレームを選び，このときの頭部の姿勢を基準の姿勢とする．頭部画像から基準姿勢に対する顔の向きを次のように測定する．

頭部画像から肌色領域を検出し，面積最大の連結領域を顔領域として，重心と外接円の半径を求める．得られた重心と半径を球体モデルの中心と半径とする．目と口の重心および左右上下端点からなる計15個の特徴点を，球体モデルに逆投影を行う．その結果，頭部重心を始点，球体に張り付いた特徴点を終点とする3次元ベクトルが得られる．顔が正面を向いているフレームの顔特徴点3次元ベクトルを基準に，頭部画像から得られる顔特徴点3次元ベクトルの向きの変化を測定する．この測定は剛体の回転運動の測定であり，四元数に基づく計算法[6]を利用した．

5. 動作の統合

身体，手，顔の3種類のモーションキャプチャで得られた動作データを統合し全身の動作を構築する．身体モデルの手部位は指の動きを含んでいないので，手のモデルの掌と入れ替える．また，身体モデルの頭部も顔モデルの頭部と入れ替える．この入れ替えにより顔の表情から指の動きをも含む全身の動作が得られるが，親部位の入れ替えには子部位の位置・姿勢の変更を伴う．

すなわち，手のモデルの掌部位の親はカメラ座標系であるが，これを身体モデルの前腕に入れ替える．まず，身体モデルの手部位の同次変換行列の並進成分は，前腕に対する手部位の相対位置であるので，これを掌の並進成分とする．身体モデルも手のモデルも同じカメラ座標系を基準にしている．手のモデルではカメラに対する掌の姿勢は，回転行列 ${}^{c}R_{palm}$ として与えられる．一方，身体モデルではカメラに対する前腕の姿勢は ${}^{c}R_{forearm}$ として計算される．前腕に対する掌の姿勢を ${}^{forearm}R_{palm}$ とすれば，${}^{c}R_{palm} = {}^{c}R_{forearm}{}^{forearm}R_{palm}$ であるので，${}^{forearm}R_{palm} = {}^{forearm}R_{c}{}^{c}R_{palm}$ として計算される．同様に顔モデルの頭部についても，その親を身体モデルの首として同次変換行列を設定する．このようにしておけば，あらゆる部位の位置姿勢を式(1)のように計算することができ，アニメーションの制作に役立つ．

6. 動作の測定実験

対象とする動作を演奏活動とした．楽器の演奏には身体の動きに加えて指の動きや顔の表情も重要と考えられるからである．今回は，良く知られた曲目の”Fly Me to the Moon”を演奏しているボーカル・ギター・ドラムスの3名の動作を測定した．

縦長のハイビジョン映像で撮影しているため，3名同時に撮影はできず，同じ曲を3名個別に演奏してもらいそれぞれ撮影した．解析対象となる映像の長さは約30秒，900フレームである．**図3**に3名のハイビジョン映像から1フレームを示す．

身体と両手の追跡結果をモデルをワイヤーフレームで画像に重ねて**図4**に示す．この追跡では，初期姿勢の付与とドリフト除去のため手動によるモデル照合が必要である．一般に動きが激しく複雑であるほど手動照合の回数が多く

図 3　測定対象
Subjects for motion capture.

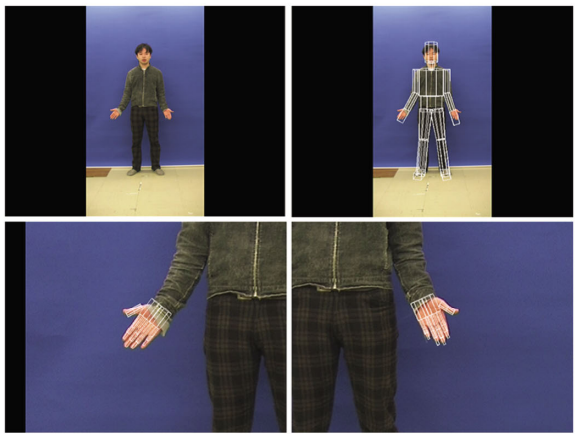

図 4　身体と手の追跡
Tracking of body and hands.

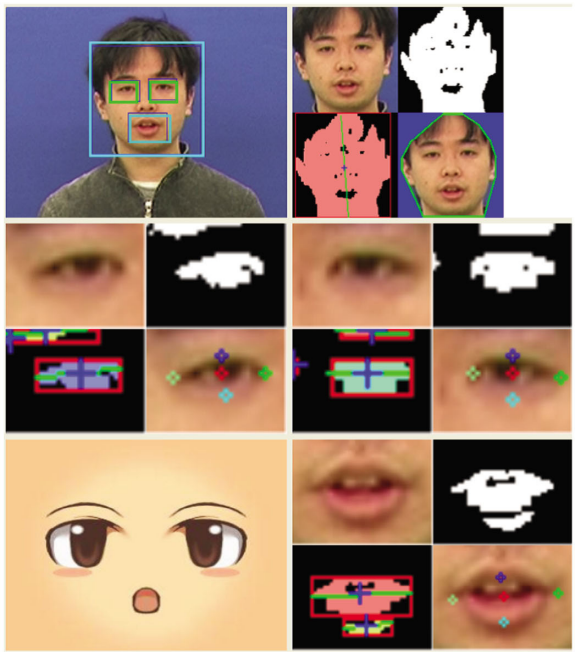

図 5　顔の表情認識過程
Detection of face features, and classification for the facial expression.

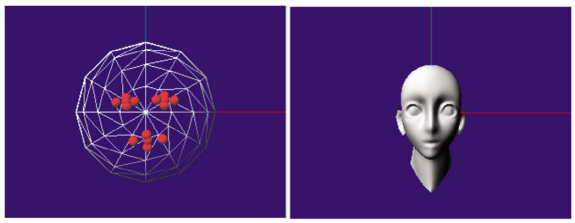

図 6　顔の向き
Face pose.

なる．身体の追跡では，平均して 36 フレームごとに手動照合を行った．手・指の動きはさらに複雑で，ドラムス，ギター，ボーカル映像ではそれぞれ，4.5，12.4，4.6 フレームごとに手動照合を行った．

顔の表情測定過程を**図 5** および **6** に示す．図 5 には 3 行 2 列の 6 枚の画像が示されている．左列上段の画像には HaarLike 特徴を使った顔，両目および口の各画像の検出結果が，部位を囲んだ矩形で示されている．右列上段の画像はさらに上下左右 4 枚の小画像を含み，顔画像の処理過程を示している．左上は顔画像，右上は閾値処理により検出された顔領域，左下は顔領域の重心と慣性主軸，右下は顔領域の凸包である．中段の左右の画像および下段の右画像は，それぞれ右目，左目，口画像の処理過程を示している．目や口画像では閾値処理で複数の領域が検出されることがある．この画像では目領域以外に眉毛領域も検出されている．このような場合は，面積最大の領域を目領域，口領域とする．右下の小画像には，抽出された目や口の左右上下の端点および中心点が示されている．左列下段には目や口の開閉を識別した結果をキャラクタの顔で示している．このキャラクタは両目口共に開いていることを表している．

図 6 左は両目と口の特徴点を頭部球体モデル上に逆投影した結果である．これらの特徴点から計算された顔の向きが，図 6 右に 3 次元顔モデルで示されている．

OpenCV に収録されている HaarLike 特徴は，正面を向いている顔の特徴であるため，顔が正面以外を向くと目や口の検出率が低下する．**図 7** には 2×2 の 4 枚の検出失敗例が示されている．左上の画像では顔が左を向いているため，左目画像の検出に失敗している．目や口の画像が正しく得られたとしても，閾値の与え方によっては目や口の領域が正しく検出できない場合もある．例えば，図 7 の左下の画像では目と眉毛が一体化している．また，右上の画像では逆に口が上唇と下唇に分かれている．このため，口が開いているにも関わらず閉じていると判断される．また，目や口の領域が正しく抽出されたとしても，他により広い領域が検出された場合には，広い領域が部位として認識される．図 7 の右下の画像では右目が正しく抽出されているにも関わらず，眉毛が目として認識された．

目や口部位の検出率は，ボーカル映像とギター映像では 91 %，ドラムス映像では 80 %であった．ドラムス映像の検出率が低いのは，顔を俯けることが多く顔の明るさが暗くなり，口と顎や目と眉毛が一体化して検出されたためである．顔の表情検出結果と身体，手・指の動作測定結果を Web 上 [7] に公開しておく．

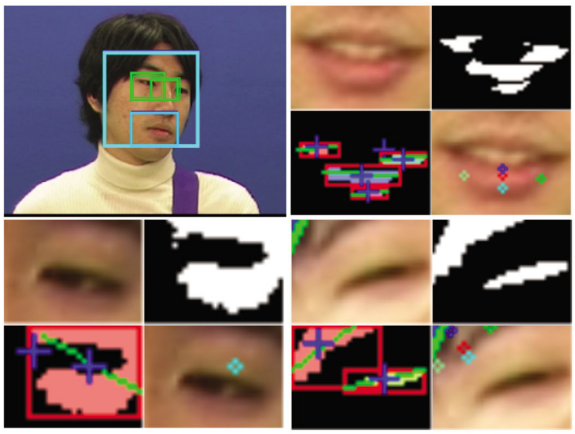

図 7　検出失敗例
Examples of detection errors.

提案手法では身体，手，顔のすべてを VGA 画像で解析している．すべてフルハイビジョン (HD) 映像で解析を行ったとした場合に対して，処理の効率化について考察しておく．手と顔は VGA 画像であるが HD 映像と解像度は同じなので，処理量は変わらない．身体の動画像は縦長の HD 映像を縦横 4 分の 1 に縮小して作られた映像なので，エッジ検出や画像差分などの処理量は 16 分の 1 となる．一方，HD 映像は VGA 映像よりも動きの測定精度が向上することが期待できるが，僅かであるが必ず含まれる測定誤差によりドリフトが生じる．精度の向上によりドリフト解消のための計算が不要になるわけではない．

7. むすび

ハイビジョン映像から，腕や脚などの身体から手の指や顔の表情に至るまで全身の動作の測定を行った．提案手法はモデルの照合や閾値の調整など手動に頼るところも多い．今後は自動化を推進すると共に，スーパーハイビジョンなどのさらに解像度の高い映像を使い，俳優の演技の詳細な測定に挑戦したい．

〔文　献〕

1) Eng Hui Loke and M. Yamamoto, "An Active Multi-Camera Motion Capture for Face, Fingers and Body", Proc. ACCV'07, pp.430-441, Tokyo (2007)
2) 山本正信, "ドリフト修正機能を有する動画像からの身体動作推定法", 信学論, **J88-D-II**, **7**, pp.1153-1165 (2005)
3) J.R. Bergen, P. Anandan, K.J. Hanna, and R. Hingorani. "Hierarchical Model-based Motion Estimation". Proc. ECCV'92, pp.237–252 (1992)
4) X. Wu, J. Lai, and X. Chen. "Efficient Human Action Detection: A Coarse-to-Fine Strategy". Proc. ICIP'10, pp.701–703 (2010)
5) R. Lienhart and J. Maydt. "An Extended Set of Haar-like Features for Rapid Object Detection", Proc. ICIP'02, **1**, pp. 900-903 (Sep. 2002)
6) B.K.P.Horn, "Closed-form Solution of Absolute Orientation Using Unit Quaternions", J. J. Opt. Soc. Am.-A, **4**, pp.629-642 (1987)
7) http://www.vision.ie.niigata-u.ac.jp/motion-capture.html

渡部　直人（わたなべ　なおと）2008 年，新潟大学工学部情報工学科卒業．2010 年，新潟大学大学院博士前期課程修了．同年，NTT コムウェア（株）入社．在学中はコンピュータグラフィックス，モーションキャプチャの研究に従事．

糸雅　亮祐（いとが　りょうすけ）2011 年，新潟大学工学部情報工学科卒業．現在，新潟大学大学院博士前期課程在学中．コンピュータビジョンの研究に従事．

佐藤　真悟（さとう　しんご）2011 年，新潟大学工学部情報工学科卒業．現在，新潟大学大学院博士前期課程在学中．モーションキャプチャによる熟練技能の解明．

三浦　裕樹（みうら　ゆうき）2011 年，新潟大学工学部情報工学科卒業．現在，新潟大学大学院博士前期課程在学中．GPGPU のコンピュータビジョンへの応用に従事．

山本　正信（やまもと　まさのぶ）1973 年，九州工業大学工学部制御工学科卒業．1975 年，東京工業大学大学院理工学研究科制御工学専攻修士課程修了．同年，電子技術総合研究所（現産総研）入所．コンピュータビジョン等の研究に従事．1989～90 年，カナダ国立研究協議会招聘研究員．現在，新潟大学工学部情報工学科教授．工博．正会員．