インバースアニメーション：映像からの動作の計測・認識・再利用

山 本 正 信†

映像からの動作測定法には 2 つの大きな利点がある．それは，身体に接触することなく自然な動作を測定することができることと，過去の人物でもその映像が残ってさえいれば，その動作を測定し再現できることである．本論文では，映像から動作を測定し動作を認識し，さらに映像制作などで動作を再利用することのできる一貫した方式を提案する．本方式では動作を表現し記述するために多関節モデルを使用する．動作の認識とアニメーションとで共通の動作モデルを使用しているので，互いに蓄積したデータを利用することができる．動作データの蓄積が多くなれば，信頼性のある認識と豊かな映像表現が期待できる．本論文で提案した方式をインバースアニメーションと呼ぶことにする．この呼称はリアルなレンダリングのために，画像からその生成過程を推定する手法がインバースレンダリングと総称されていることにヒントを得た．

Inverse Animation: An Approach to Image-based Motion Caputuring, Gesture Recognition and Performance Animation

Masanobu Yamamoto†

An image-based motion capturing has two major advantages. First, it can measure natural movement of human body non-invasively. Second, for even a person passed away, as far as his/her action is recorded on a film or video, it can reconstruct his/her movement. This paper proposes a novel approach to motion capturing, gesture recognition and creating performance animation. This approach uses an articulated model to describe a human body in motion. The common model makes it possible for gesture recognition and performance animation to share each other's action data. The richer the action databases are, one can expect the more reliable gesture recognition and the richer performance in animation. We shall call the proposed approach an inverse animation. This naming is inspired by inverse rendering which is technique to estimate a process generating an image.

1. はじめに

普段何気なく見ている人間の動作ではあるが，動作を測定すればコンピュータアニメーションなどで再利用することが可能である．また，近年セキュリティへの関心が高まるにつれ，人間の動作を認識し動作の意味や意図を理解する必要性が増大してきている．

人間の動作を測定しそれをコンピュータグラフィックス上で再現するシステムはモーションキャプチャと呼ばれている．これまでのモーションキャプチャは，身体にあらかじめデバイスやマーカなどを取り付けた方式が多い．これらを接触型モーションキャプチャ[18]と呼ぶ．接触型はデバイスが身体の自由度を損なったり，測定対象者に機材やマーカを意識させたりして，自然な動作を測定することが難しい．これに対し，身体にデバイスやマーカなどをいっさい取り付けず，映像のみから動作を測定しようとする方式は非接触型となる．

映像からの動作測定法には 2 つの大きな利点がある．その 1 つは，対象者に意識されることなく自然な動作が測定できることである．これは，セキュリティを目的とした動作認識にとって重要な利点である．もう 1 つは，過去の人物でもその映像が残ってさえいれば，その動作を測定し再現できることである．これは，映画の発明以来蓄積されてきたすべての映像から人間の動作が測定可能であることを意味し，アニメーションなどで再利用できる動作データが飛躍的に増大することが期待される．

本論文では，映像から動作を測定し動作を認識し，さらに映像制作などで動作を再利用することのできる一貫した方式を提案する．

本方式では動作を表現し記述するために多関節モデルを使用する．このモデルは動作が身体固有の表現法

† 新潟大学工学部情報工学科
Department of Information Engineering, Niigata University

で記述されているため，カメラの視点などの撮影条件に依存しない認識系を構築することができる．また，動作の認識とアニメーションとで共通の動作モデルを使用しているので，互いに蓄積したデータを利用することができる．動作データの蓄積が多くなれば，信頼性のある認識と豊かな映像表現が期待できる．

近年，リアルな画像を作り出すために，コンピュータビジョンとコンピュータグラフィックスは互いに依存性を高めている．レンダリングのために，画像からその生成過程を推定する手法をインバースレンダリング（inverse rendering）と総称している．これに対し，リアルなアニメーションのために，映像から身体の動作など映像の生成過程を推定する手法をインバースアニメーション（inverse animation）と呼ぶことにする．もちろん，物体の運動の解析は動画像解析として古くから行われてきた．インバースアニメーションとは，映像の制作や動作の認識など応用先を強調した呼称でもある．本論文では，インバースアニメーションの一手法を提案していることになる．

動作を測定する準備として，2 章では身体多関節モデルについて述べる．3 章では，身体モデルと映像との照合によって動作を測定する方法を示す[32),37)]．4 章では，動作データを使った動作認識の手法を示す[19),20)]．5 章では，動作の再利用例として映画アセスメントシステムなどについて述べる[8),35),36)]．6 章では，論文の総括と今後の課題について述べる．

2. 身体モデル

身体の多関節モデルの例を**図 1** 左に示す．このモデルは各部位が関節でつながった木構造で表されている．各部位の連接関係を図 1 右に示す．矢印の向きは親から子への向きを表している．部位の運動はそのすべての親の運動が影響するが，それ以外の部位の運動は影響しない．部位間の親子関係を遡れば，最上位の親から親子順に連鎖番号を付けることができる．それぞれの連鎖番号には実際の部位が対応している．たとえば，下腕の運動を求めるときは，腰部，胸部，上腕，下腕の順に，1，2，3，4 と番号を付ける．

各部位は固有の座標系を持っている．部位座標系は通常原点を関節位置（腰部や胸部では重心位置）に置き，座標軸の 1 つを部位の体軸に一致させておく．

身体の位置・姿勢・運動は部位座標系の位置・姿勢・運動で表すことができる．身体各部位の姿勢は，その親の座標系を基準に表す．すなわち，部位 i の座標系 Σ_i は，その親である部位 $i-1$ の座標系 Σ_{i-1} を基準に並進と回転で表す．ただし，最上位の腰部の親

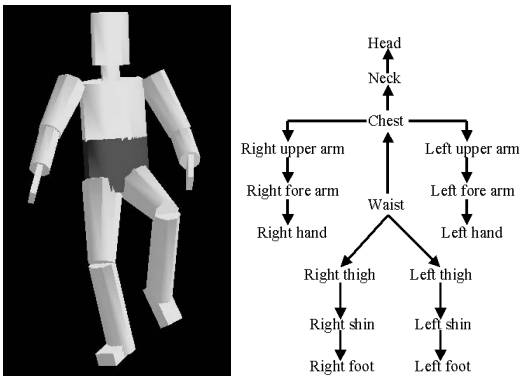


図 1 身体の多関節モデル
Fig. 1 An articulated model for human body.

はカメラ座標系とする．子から親座標系の変換を同次変換行列で

$$\tilde{T}_i = \begin{pmatrix} T_i(\theta) & \mathbf{t}_i \\ \mathbf{0}^\top & 1 \end{pmatrix}$$

と表す．ここで，$T_i(\theta)$，$(\theta = (\theta_{z_i}, \theta_{y_i}, \theta_{x_i}))$ は回転行列で，そのオイラー分解を z，y，x 軸回りの回転角 θ_{z_i}，θ_{y_i}，θ_{x_i} で表す．また，$\mathbf{t}_i$ は並進ベクトル $(t_{x_i}, t_{y_i}, t_{z_i})$ である．

カメラ座標系で表された点 $\mathbf{p}$ が，部位 i の座標系 Σ_i では $\mathbf{p}_i$ で表されているとする．それぞれの点の同次座標系表示を，$\tilde{\mathbf{p}}_i = (\mathbf{p}_i, 1)^\top$，$\tilde{\mathbf{p}} = (\mathbf{p}, 1)^\top$ とする．

このとき部位座標系からカメラ座標系への変換は，

$$\tilde{\mathbf{p}} = \tilde{T}_1 \tilde{T}_2 \cdots \tilde{T}_i \tilde{\mathbf{p}}_i \qquad (1)$$

となる．

部位 i がさらに自身の座標系を基準に並進・回転運動を行ったとする．この運動を同次変換行列で

$$\tilde{R}_i = \begin{pmatrix} R_i(\phi) & \mathbf{r}_i \\ \mathbf{0}^\top & 1 \end{pmatrix}$$

と表す．ここで，$R_i(\phi)$，$(\phi = (\phi_{z_i}, \phi_{y_i}, \phi_{x_i}))$ は回転行列で，そのオイラー分解を z，y，x 軸回りの回転角 ϕ_{z_i}，ϕ_{y_i}，ϕ_{x_i} で表す．また，$\mathbf{r}_i$ は並進ベクトル $(r_{x_i}, r_{y_i}, r_{z_i})$ である．

各部位は親の部位とリンク結合しているため，自由度は回転のみである．しかし，最上位の腰部は並進を含めた 6 自由度である．

モデルが運動したとき，位置 $\tilde{\mathbf{p}}_i$ はカメラ座標系では $\tilde{\mathbf{p}}'$ に移動したとする．この変換は，

$$\tilde{\mathbf{p}}' = \tilde{T}_1 \tilde{R}_1 \tilde{T}_2 \tilde{R}_2 \cdots \tilde{T}_i \tilde{R}_i \tilde{\mathbf{p}}_i \qquad (2)$$

で与えられる．

部位の回転運動は姿勢角 θ の変位としても表すことができる．しかし，姿勢をオイラー角で表している限

り y 軸回りの回転が $\pm\pi/2$ のときジンバルロック☆に陥ることが知られている[5]．ジンバルロックを回避するためには，動きを動座標系あるいは四元数で表すなどの方法があるが，ここでは動座標系で表している．

3. 動作計測

身体モデルを画面上の身体像にあてはめたとき，モデルから身体の位置・姿勢を知ることができる[1),2),7),9),13),22),23),25),29)]．フレーム順にモデルをあてはめれば動作が測定できる．一方，位置・姿勢の変位分は連続する画像間の差分値を使って容易に推定できる[3),12),33),34)]．したがって，動画像の最初のフレームでモデルをあてはめ，初期位置・姿勢を得た後，これに姿勢の変位を累積しても動作を知ることができる．1 フレームごとにモデルを身体像にあてはめるよりは計算量が少なくてすむ．

ところが，このアプローチには 2 つの大きな問題点がある．1 つは画像から身体を抽出しモデルを 3 次元的にあてはめる課題が，これまでのところ必ずしも容易ではないことである．たとえば，身体の抽出では，背景差分や画像間差分がよく使われるが，カメラが固定されていることが前提である．肌や衣服の色などの特徴を利用すれば，カメラの固定は必要ではないが，背景にも同様な特徴が存在する場合には抽出された領域を区別する必要がある．また，モデルのあてはめでは，画像上で部位が隠されているときにはその姿勢を推測することは難しい．

したがって，現状では，この課題のすべてを自動的に行うのではなく，ところどころ人の支援を仰ぐのが現実的な解決法である．人間にとって身体の部位を認識することは容易であり，隠された部位の姿勢の推測も自己の体験から可能と思われる．そこで，**図 2** のようなモデルの位置・姿勢を与えることのできる簡単な GUI を用意した．モデルの 3 次元姿勢を右の 3 面図で確認しながら，左の大画面に投影されたモデルのあてはめを手作業で行うことにする．幸いにして，アニメーション業界では，キーフレームの設定において，人がキャラクタの姿勢を設定するのは慣れていて苦にしていない．

もう 1 つの問題は，位置・姿勢の変位分の推定時に起こる．変位の累積から得た姿勢は実際の姿勢から時間とともに大きくずれてしまう．このずれをドリフトと呼んでいる．ドリフトは推定された変位の誤差がわずかでも，それに一定の傾向があるならば累積されることによって起こる．たとえば，**図 3** 左端のように静止画を縦横方向に少しずつ移動させ，図の黒線に示すような円弧運動で一周させたとする．図の矩形領域でオプティカルフローを求め矩形の中心位置に逐次累積させると，図 3 中の白線円弧のように，1 周したところでずれを生じる．これがドリフトである．この例では，矩形内に他の動きの混入はないが，身体の動作測定ではモデルが身体からずれると測定領域内に他の運動が含まれ，それがますますドリフトを増大させる．

本論文では簡単で強力なドリフト修正法を示す[37)]．動画像の最後のフレームでもモデルをあてはめ，得られた最終位置姿勢に対し，累積により得た位置姿勢が一致するように位置姿勢の修正を行う．それには，最終フレームでのドリフトからフレーム間の平均ドリフトを求める．フレーム順に，得られた身体位置姿勢から平均ドリフトを差し引くことにより位置姿勢の修正を行う．図 3 右端は矩形中心位置の修正結果が白線で示されている．

位置の修正はこのように容易であるが，姿勢の修正

図 2 モデルフィッティングのための GUI
Fig. 2 A GUI for model fitting.

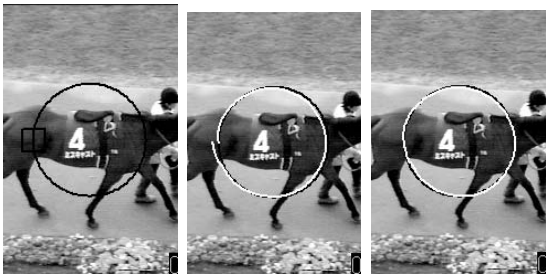


図 3 ドリフトの発生と修正
Fig. 3 Drift and correction.

☆ y 軸回りに $\pm\pi/2$ 回転させたとき，x 軸の方向が最初の z 軸の方向と一致し，2 軸回りの回転が 1 軸回りの回転に縮退する現象．

は技巧を要する．1 つの身体部位について，開始と終了キーフレームの姿勢をそれぞれ T_0，T_n とする．また，測定されたフレーム間動作を順に $R_1, R_2, \cdots, R_n$ とする．ここで，姿勢，動作ともに直交行列で表している．動作の累積による最終姿勢は，$T_0R_1R_2\cdots R_n$ であるが，ドリフトのため T_n に等しくない．フレームごとの平均ドリフトの補正を X とする．この補正量は，

$$T_0R_1XR_2X\cdots R_nX = T_n \qquad (3)$$

を解くことによって得られる．式 (3) は X に関する高次方程式であるが，補正行列 X が微小であるとき，式 (3) は 3 つの回転運動パラメータに関する線形連立方程式で近似できる．

実際，X は 3 自由度の回転行列なので，その回転運動パラメータを 3 次元ベクトル $\delta\mathbf{x}$ とおく．式 (3) の左辺は，$F(\delta\mathbf{x}) = T_0R_1XR_2X\cdots R_nX$ と書ける．$F(\delta\mathbf{x})$ を $\delta\mathbf{x} = \mathbf{0}$ でテーラー展開し，2 次以上の微小項を無視すれば，

$$F(\delta\mathbf{x}) \approx F(\mathbf{0}) + \left(\frac{\partial F(\delta\mathbf{x})}{\partial \delta\mathbf{x}}\right)^{\top} \delta\mathbf{x}$$

を得る．この式が式 (3) の右辺 T_n に等しいとすれば，$\delta\mathbf{x}$ に関する線形連立方程式が得られ，補正 X はこの 3 元 1 次連立方程式を解くだけで得られる．ただし，近似解であるので正しい解はニュートン法による繰返しが必要である☆．

実際は位置・姿勢が補正されてもモデルは必ずしも身体に一致するとは限らない．補正された位置姿勢で再度フレーム間動作を測定し，初期位置・姿勢に累積させる．ここで，ドリフトが発生すれば再度補正を行う．動作の測定とドリフトの補正を交互に繰り返すことによりモデルは身体に一致する．実験によればこの繰返しは数回で十分である．

ただし，長時間の動作追跡では，位置・姿勢を開始・終了フレームで与えるだけでは不十分で，いくつかの中間フレームでも，位置・姿勢を与えることになる．位置・姿勢が与えられたフレームをキーフレームと呼ぶ．

図 4 は野球の投球動作を追跡した結果を，モデルを画像に重ねて表示している．全部で 188 フレームである．モデルを照合させるキーフレームは 16 個であった．**図 5** はこの投球動作をバッターの視点から再現した CG 映像である．

ドリフトの修正法は，早い時期に大田ら[21] が提案している．しかし，彼らの手法は最終姿勢を弛緩法により中間フレームへ伝播させることによって姿勢の修正を行うため収束するまで多くの計算量を必要とする．

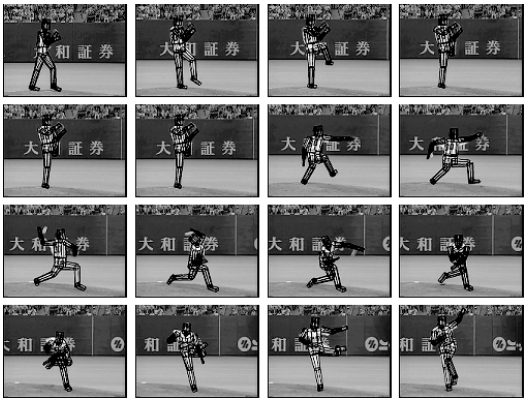


図 4 投球動作の追跡結果
Fig. 4 Tracking result of pitching in the baseball.

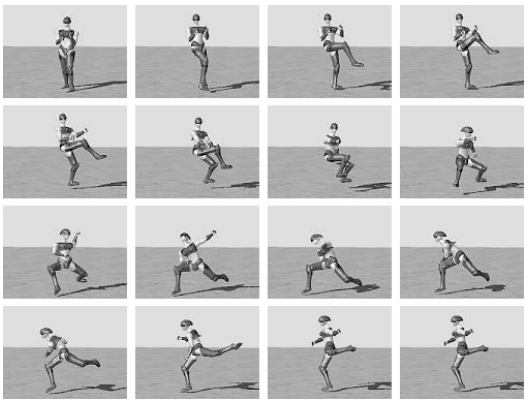


図 5 投球動作の CG 映像．追跡結果からの再構成
Fig. 5 Performance animation from capturing of pitching.

さて，単眼カメラからの映像を使用している限りさらに 2 つの問題点がある．1 つは隠れであり，もう 1 つは奥行き方向に推定誤差を生じることである．

隠れが一時的なら，隠れた部位の姿勢は隠れの前後の姿勢からロボットの作業計画法などを使って補間することができる[30),37)]．まったく観測できなかった部位の扱いは 5 章に示す．

奥行き方向の動きの推定誤差は，環境からの束縛を利用して抑えることもできる．身体はその動作環境から様々な束縛を受けている．たとえば，スキーやスケートをしているとき，その足先は滑走面に拘束されている．これらの環境からの束縛を利用することにより，動作の自由度を制限し誤差の発生を抑えることができる[32)]．**図 6** は滑降しているスキー動作の追跡結果である．滑走面の拘束を利用した場合（図中）には，利用しなかった場合（図右端）に比べて，両足先が地面に接地している．

☆ 計算方法の詳細は文献 37) を参照.

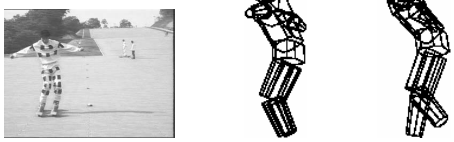


図 6 環境拘束を利用した奥行き誤差の抑制
Fig. 6 Pose correction by scene constraints.

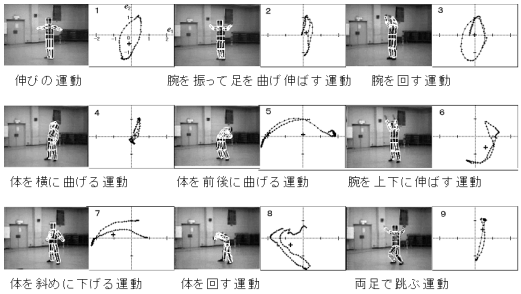


図 7 ラジオ体操の運動と固有平面上への動作軌道
Fig. 7 Gesture on the eigenplane in the radio gymnastic exercises.

4. 動作の認識

動作データは時系列データであるので，同じく時系列データを扱っている音声認識の手法が利用できる．実際，隠れマルコフモデル（HMM）[38] や動的計画法（DTW）[28] などが動作認識に有用であることが示されている．

一方，動作データは動作パラメータ空間中の軌道として表される．このパラメータ空間の次元を 2 次元に圧縮すれば，軌道は平面上の線画として描くこともできる．線画ならばその認識は文字認識の問題である．また，動作を基本動作のつながり，すなわち記号列で表すことができれば，動作の認識は言語理解の手法で扱える．本論文では，文字認識の手法や言語理解の手法が動作認識に使えることを示す．

4.1 文字認識手法の適用

動作中の姿勢データは，パラメータ空間中に 1 つの軌道を描く．同じ種類の動作であればその軌道はほぼ同じ軌道を描く．一方，動作の種類が異なれば別の軌道を描くことになる．したがって，描かれた軌道の特徴から動作の識別を行うことが可能となる[4]．

しかしながら，人体の姿勢を表すには多数のパラメータを必要とする．たとえば，図 1 の人体モデルは 30 個のパラメータを使用している．パラメータの数が多くなれば，識別のための計算量も多くなる．また，高次元の軌道を分かりやすく表示することも難しくなる．したがって，少数のパラメータを使って，動作の特徴を失わずに表現できることが望ましい．

様々な動作の姿勢データ列を KL 展開したとき，データのばらつきが大きな順に新たな座標軸が得られる．そして，各座標軸に対応する新しいパラメータで姿勢を表すことができる．データのばらつきの大きなパラメータ値は，姿勢の変化に敏感であり独立性が高いといえる．その逆に，ばらつきが小さいパラメータ値は姿勢の変化に鈍感であり，姿勢を表すにはなくてもよい．これは，身体の各部は互いに関連して動いているため，独立な姿勢パラメータの数は少ないと考えられるからである．したがって，データのばらつきの大きな順に数本の座標軸を選び，これらの座標軸から構成される小さなパラメータ空間にも十分に動作の特徴は含まれている．特に，この小パラメータ空間が 2 次元の場合を固有平面と呼ぶ．

固有平面上へ射影された姿勢パラメータ列も動作ごとに一定の軌道を描く．この軌道は平面上に描かれた文字と見なすこともできる．したがって，動作の認識問題は文字認識の問題と等価になる[20]．

このことをラジオ体操を例に示す．ラジオ体操第 1 に含まれる 9 種類の動作について，それらの動作中の 1 コマが**図 7** に示されている．その右には，各動作の軌跡が固有平面上に描かれている．これらの図から，動作の種類が異なれば描く軌跡も異なることが分かる．

4.2 言語処理手法の適用

人間の行動はそのほとんどが習慣的な行動である．習慣的な行動は「一連のプログラム化された半自動的な諸行動の体系」[24] と見なすことができる．これは，行動は短い基本動作の組合せから構成されることを意味する．行動が基本動作列で表すことができれば，記号列からの動作認識は言語理解の問題と等価になる．和田ら[31] は，非決定性有限オートマトンで基本動作の遷移を記述し，この知識を使って頑強な動作認識システムを構成している．

動作の規則を文法として表すことができれば，その文法を使って動作の認識が可能である．動作の規則は観察により知ることができるが，マニュアルにより動作が定められている場合には動作マニュアルが規則集になる．動作マニュアルのある例としては，ダンスの振り付け，茶道の点前，自動車や航空機などの運転操作がある．このうち，茶道の点前の例を示す．

茶道における点前とは，湯を沸かし，茶を点て，それを飲む動作であるが，茶道の長い歴史の中でその作法が確立されている．客や季節，使用する茶道具などに応じて様々な種類の点前が提供されている．代表的

な点前は 30 種類ともいわれ，それぞれ約 200 個の基本動作から構成されている．

点前は動作プランニングの一種であり，階層構造で表すことができる．最上位は点前の種類を表し，点前は準備，喫茶，仕舞の部分動作から構成され，それぞれ具体的な動作，さらには基本動作へと展開される．このような階層構造は自然言語の構造に似ており，そのため点前を記述するためには文脈自由文法が有効であると考えられる．

初歩的な点前である薄茶点前と濃茶点前の一部を，茶道の教本[27] から抜粋し**表 1** に示す．濃茶点前はステップ 8 で，基本動作「主客総礼をする」（亭主が客に向かってお辞儀をする）があるが，薄茶点前ではない．その他の基本動作は共通する．この部分に注目して最も簡単な動作認識例を示す．

記号で「柄杓を置く」動作を h,「建水を進める」動作を k,「総礼をする」動作を r，とすれば，薄茶点前は hk，濃茶点前は hrk となる．この 2 つの記号列を認識（導出）する文脈自由文法を**表 2** に示す．ここで，終端記号を $\{h, r, k\}$，非終端記号を $\{A\}$，開始記号を $\{S\}$ とする．

この文法は 2 つの記号列を認識することができる．では，両者を見分けるにはどうすればよいであろうか．まず，この表 2 の右半分に示すように，点前ごとに生成規則に確率を与え確率文脈自由文法にする．点前 hrk は生成規則 3 と 2 から導出されるが，この点前が薄茶点前である確率は，導出に使った規則の生成確率の積で与えられる．この場合は確率は $0 \times 1 = 0$ となる．一方，濃茶点前である確率は $1 \times 1 = 1$ となる．点前 hk では，これを導出する規則は 1 であるので，薄茶である確率は 1，濃茶である確率は 0 となる．規則の生成確率を変えるだけで，それぞれの動作を解釈したときの確率が計算できる．そして，確率の最も高い解釈を認識結果とすればよい．

文法による動作認識では，姿勢データ列が基本動作列に記号化できることが前提条件となっている．変換方法としては，基本動作の境目を検出し，検出された区間がどの基本動作に対応するかを決定する．

しかし，この記号化は必ずしも正確ではないので，誤った基本動作列からでも動作が認識できる必要がある．Ivanov ら[11] は，この問題をリアルタイムパーサを使って解決している．CYK などの従来の構文解析機は，記号列全体が与えられることが前提であるが，Earley-Stolcke の構文解析機は，記号列が一部ずつ与えられたとしても，予測と誤りの修正を逐次行うことができる．

しかしながら，いったん記号化した後では記号化の誤りを完全に修復させることは難しい．我々[19] は，動作を加速度を使って区分けした後，区間の基本動作への対応付けは保留し，対応可能なあらゆる基本動作列を残しておく．構文解析機により導出可能な基本動作列を絞り，残った候補から最も導出確率が高くなる解釈を認識結果とする．この手法はベイズの意味で最良の認識結果を与えるが，動作を区分けした区間の数が多くなると対応可能な基本動作列数が大きくなるのが欠点である．

表 1 薄茶点前と濃茶点前の動作比較
Table 1 A context-free grammar of temaes.

#	薄茶点前	濃茶点前
・	…	…
5	蓋置きを敷板の左隅に置く	蓋置きを風呂敷板の左隅に置く
6	柄杓を右手に持ち直す	柄杓を右手に持ち直す
7	柄杓を蓋置きの上に合をのせ引く	柄杓を蓋置きの上に合をのせ引く
8		主客総礼をする
9	左手で建水を膝前の線まで進める	左手で建水を膝前の線まで進める
10	右手で茶碗をとり膝前中央の向こうに置く	右手で茶碗をとり膝前中央の向こうに置く
・	…	…

表 2 点前の文脈自由文法
Table 2 A comparison of actions between Usucha and Koicha temaes.

生成規則					薄茶点前	濃茶点前
1	S	→	h	k	1	0
2	S	→	A	k	0	1
3	A	→	h	r	1	1

5. 動作の再利用

映画やドラマのカメラワークでは，ミディアムショットやクローズアップなどが多用され，フルショットで演技が撮影されることが少ない．そのため，これらの映像からは部分的な動作しか得られない．これは 3 章で扱った一時的な隠れとは異なる．観測されなかった動作を復元し全身の動作データを作っておく．

動作データを使って制作される作品はパフォーマンスアニメーションと呼ばれている．本章では，パフォーマンスアニメーションの活用法として，映画の制作に入る前に，作品のできばえを事前に評価するシステムを示す．また，これまでは映像を制作した後で音声や音楽を加えていたが，ここでは音声や音楽から動作映像を自動的に生成する手法を示す．

5.1 隠れた動作の復元

全身の動作を得るためには測定された部位の動作と

マッチした動作を復元しなくてはならない．既存の動作データを利用した動作の復元法を 2 つ示す[36]．1 つは固有空間法でありもう 1 つは部分照合法である．

固有空間法

復元対象と同じ動作を演じたときのサンプル姿勢列 $\mathbf{s}_1, \mathbf{s}_2, \cdots$ を用意する．この姿勢データを KL 展開し，元の姿勢パラメータ空間から次元の小さな固有空間への変換行列を E とする．元の空間の姿勢 $\mathbf{x}$ から固有空間上の姿勢 $\mathbf{y}$ への変換を $\mathbf{y} = E^\top \mathbf{x}$ とする．逆変換は次式で近似できる．

$$\mathbf{x} \simeq E\mathbf{y} \tag{4}$$

今，姿勢ベクトルの中で最初から k 番目の姿勢までが測定され，残りが隠れによって未知だとしよう．測定された姿勢ベクトルを $\tilde{\mathbf{x}}$，未知の姿勢ベクトルを $\mathbf{x}'$ とし，逆変換 E の上位 k 行を $\tilde{E}$，$k+1$ 行目以降を E' とすれば，式 (4) は，

$$\begin{pmatrix} \tilde{\mathbf{x}} \\ \mathbf{x}' \end{pmatrix} = \begin{pmatrix} \tilde{E} \\ E' \end{pmatrix} \mathbf{y} \tag{5}$$

と書ける．さらに，上式は 2 つの線形方程式となる．

$$\tilde{\mathbf{x}} = \tilde{E}\mathbf{y} \tag{6}$$

$$\mathbf{x}' = E'\mathbf{y} \tag{7}$$

式 (6) を $\mathbf{y}$ について解き，得られた $\mathbf{y}$ を式 (7) に代入すれば未知の姿勢 $\mathbf{x}'$ が得られる．

部分照合法

画像上で観測できた部分 $\tilde{\mathbf{x}}$ とモデル上の対応部分 $\tilde{\mathbf{s}}_i$ との差異の二乗和を最小にする解を求める．すなわち，

$$j = \arg\min_i \|\tilde{\mathbf{x}} - \tilde{\mathbf{s}}_i\|^2 \tag{8}$$

より，身体の姿勢に最も近いサンプル姿勢 $\mathbf{s}_j$ が得られる．

しかし，サンプル内にあまり近い姿勢がない場合には，選択された姿勢と実際の姿勢とのずれは大きく信頼度の低い姿勢となる．このときサンプル姿勢列は滑らかに推移せず，これはジッタと呼ばれている．そこで，姿勢列が滑らかに推移するように姿勢を部分的に修正する．ただし，式 (8) で得られた最小差異の逆数を信頼度とし，信頼度の大きな姿勢ほど修正が小幅であるようにする．**図 8** に実際の映画でミディアムショットで撮られた歩行時の上半身映像から脚の歩行動作を復元した結果を示す．

固有空間法は部分照合法に比べて計算量がはるかに少ない．しかし，隠された部位が多くなるほど固有空間法による復元は難しく，部分照合法の方が信頼度の高い復元が可能であることが実験的に確かめられている．この理由は次のように考えられる．固有空間上の姿勢 $\mathbf{y}$ は元の空間の姿勢ベクトル $\mathbf{x}$ の射影であるが，式 (6) では姿勢の一部 $\tilde{\mathbf{x}}$ から得ている．このとき，本来の姿勢からずれる危険性があり，このずれが姿勢の復元 (7) で増幅されることも考えられる．これに対し，部分照合法は既存の姿勢データ中から姿勢を選択するのでつねに安定した結果が得られる．

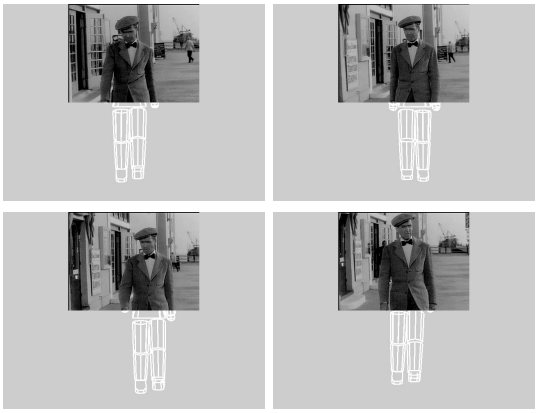


図 8 部分照合法による隠れた部位の追跡
Fig. 8 Tracking occluded parts by model matching method from a movie, namely, “Glenn Miller Story,” 1953, ©Universal Pictures Corporation.

5.2 映画制作のアセスメント

伝統的に映画製作は，非常に多くの経済的負担と時間および労働力をかけてきた．芸術的価値の高い映像を撮るためには，多大な投資もやむをえないが，撮影現場での試行錯誤を減らせば，このコストを大幅に削減することができる．本論文では，映画の撮影に入る前に，撮るべき映像を様々な角度から仮想空間上で事前に評価するシステムを示す[8]．

このシステムは，動作データと俳優のモデルを使って簡単なアニメーションを制作することができる．動作データは 3 章で提案したモーションキャプチャにより測定されたものである．現在のところ，立ち上がる，歩く，踊るなどの動作のほか，4 章の動作認識で使用した体操や茶道点前の動作も蓄積されている．

図 9 にシステムの GUI を示す．右半分のワークスペースでカメラワークと撮影リズム，映像の構図，俳優の配置と動作などを設定し，左半分のモニタで表示する．映画監督は，1 ユーザとして，全体を通した映像の構成を事前に評価することができる．

このシステムがどの程度撮影コストを節約できるかを実際の撮影で比較した．まず，数十秒のプロット☆を制作するための絵コンテを用意した．数名のカメラマ

☆ 少女が夜道を歩いているときに吸血鬼に遭遇するシーン．

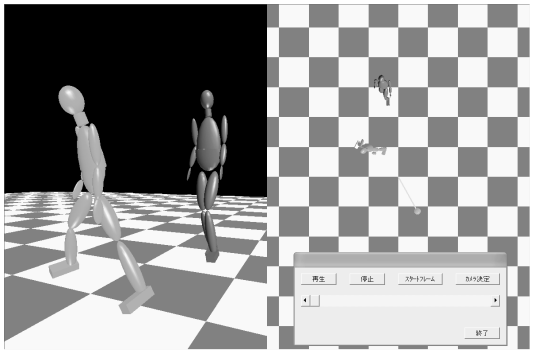


図 9 映画アセスメントシステムの GUI
Fig. 9 A GUI for shot assessment.

ン☆を 2 グループに分け，1 つのグループには絵コンテのみから撮影を行わせた．もう 1 つのグループには提案したシステムを使ってカメラワークを検討させた後，現場で撮影を行わせた．それぞれ納得のゆく映像が得られるまで撮影を繰り返させた．その結果，提案したシステムを使ったグループの撮影回数は平均 1.1 回，絵コンテのみのグループは平均 3.2 回であった．提案システムの使用により撮り直しの回数が約 3 分の 1 に減っている．事前に評価しておいた効果が現れている．なお，両グループとも最終的に決定したカメラワークはほぼ同じものであり，カメラワークの定石[14] どおりのものであった．プロのカメラマンにとっては，提案手法は新しいカメラワークを発見する助けになるものと期待される．

従来コンピューグラフィックスは映像の制作技術であったが，ここでは制作コスト削減の手段，あるいは芸術家の創造性を高める道具として使われている．

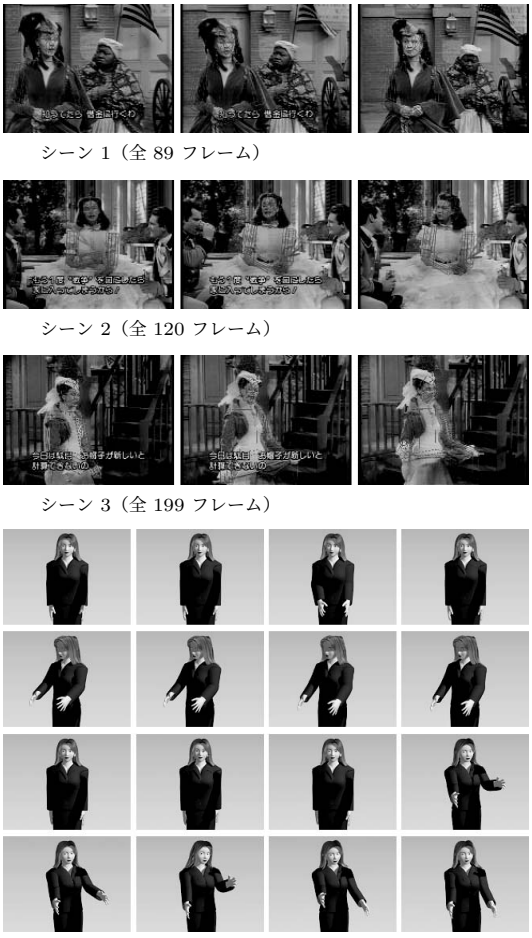
5.3 音声・音楽からの動作の生成

これまでの映像コンテンツを分析すると，ほとんどが俳優間の会話で成り立っている．したがって，映像コンテンツの制作には，映像に対応した音声，あるいは音声に対応した映像の同時制作が必要になってくる．映画やテレビドラマから動作を測定したとき，俳優の音声や音楽も同期して得ることができる．この動作や音声・音楽は互いに密接に関連している．もし，動作と音声や音楽との関係を見い出すことができれば，音声や音楽からそれにふさわしい演技を作り出すことができる．本論文では簡単な生成実験を試みたので示す[35]．

時刻 t での姿勢が，その時刻での音と過去の動作から決定されるとすれば，その関係は

$$y_t = f(x_t, y_{t-1}, y_{t-2}, \cdots) \qquad (9)$$

☆ 映像カメラマンとしては素人の工学系学生 10 名.



シーン 1（全 89 フレーム）

シーン 2（全 120 フレーム）

シーン 3（全 199 フレーム）

動作の生成映像（全 389 フレーム）

図 10 動作と音の関係の学習および音からの動作の生成
Fig. 10 Learning relation between gesture and sound from "Gone with the wind," 1939, ©MGM/UA Home Entertainment Inc. and Turner Entertainment Co., and avatar in action from sound.

と表される．ここで，y は関節角度の 1 つ，x は音の強度，f は関係を表す関数である．

この関数をテーラー展開し，音は m 次までの多項式，姿勢は線形式として次のように近似する．

$$y_t = a_0 + a_1 x_t + a_2 x_t^2 + \cdots + a_m x_t^m + b_1 y_{t-1} + b_{t-2} y_{t-2} + \cdots + b_n y_{t-n} \qquad (10)$$

係数 a_i と b_j を動作データと音データから学習する．音データのサンプリングレートは動作データのそれよりも高い．したがって，ここではフレーム間の音振幅の二乗和，すなわちパワーを音の強度とする．

映画「風と共に去りぬ」から，主演女優が演じている 3〜10 秒の長さの 4 つの映像シーンを切り出す．順にシーン 1〜4 と名づける．**図 10** の上から 3 行の映像は，それぞれシーン 1〜3 から 3 コマずつ表示した

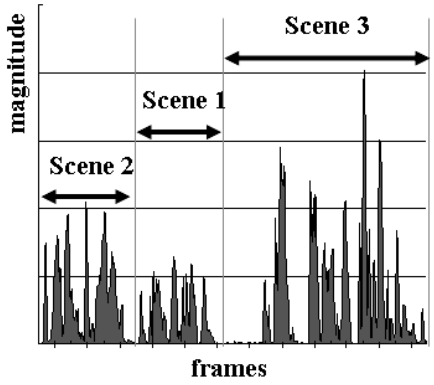


図 11 台詞音声と関連付けられた動作モデル
Fig. 11 Sound-action model related with scenario.

ものである．各シーンの映像から俳優の動作を測定するとともに，同期している音の強度データを得る．得られた動作データと音データを式 (10) にあてはめ，両者の関係を式の係数パラメータとして学習する．ここでは簡単のため式 (10) の多項式の次数を 2，姿勢に関する線形項は無視した．このようにして，シーン 1〜3 の音・動作モデルを得る．シーン 1 からは暗い表情，シーン 2 からは冷たい表情，シーン 3 からは甘えた表情で話す音・動作モデルがそれぞれ得られた．

一方，シーン 4 から音声データのみを取り出し，その音声にふさわしい動作を音・動作モデルから生成する．

この映画はシナリオ[10] が公開されている．シーン 4 の音声に対応する台詞（line）は順に次の 3 つであった．シナリオ[10] から抜粋して示す．

line1 She just got tired of waiting, was afraid she'd be an old maid.

line2 Oh, I'm sorry to be the one to tell you!

line3 Oh, it's cold and I left my muff at home. Would you — would you mind if I put my hand in your pocket? (with an intimate gesture)

図 11 にはシーン 4（389 フレーム）の音の強度が時間の推移とともに示されている．台詞の切れ目は手作業で与えた．最初の台詞は事実の伝達，2 番目の台詞は謝罪，最後の台詞は哀願であった．したがって，事実の伝達には冷たい表情（シーン 2），謝罪には暗い表情（シーン 1），哀願には甘えた表情（シーン 3）の音・動作の関係モデルを与えた．音データに対し式 (10) から姿勢を計算し，アニメーション映像を作成した．図 10 の下半分は作成した映像を示す．第 1 行と第 2 行は最初と 2 番目の台詞に対応する．それぞれ映像から 4 コマずつ抜き出して表示した．第 3〜4 行目は最後の台詞に対応した生成映像である．

音声，音楽[16]，文章[26] などからキャラクタの動作を生成する試みは数多くあるが，このような簡単な手法でも動作の生成は可能である．ただし，得られた動作の評価はこれからの課題である．

6. おわりに

映像から動作を測定し動作認識やコンピュータアニメーションで再利用する手法を示した．

動作の測定では，キーフレームでのモデル照合とキーフレーム間での動作追跡を組み合わせることが，現時点での最良の手法であろう．しかし，実際に両者を組み合わせた例は少ない．これは，画像から身体を抽出し部位を特定することが，画像認識の基本問題であり，いまだ汎用手法が存在しないことがネックとなっている．問題を簡単にするためには，対象とするシーンや動作を限ることも考えられる．

一方，動作の認識では，演じる人の動作を中心に議論してきたが，人が使用する道具や対象に着目することも考えられる．茶道では茶筅や茶碗の動き，料理では包丁の動きなどからでも動作を認識することが可能である．ただし，どの情報チャネルでもそれだけで十分であるとはいえない．マルチモーダルな情報をいかに関係付けいかに相補えるか[15] がこれからの動作認識の鍵と思われる．

動作データの再利用はアニメーションやセキュリティだけではない．医療経過の評価やスポーツにおけるスキルの評価[17] などにも利用されている．

本論文では，提案手法をインバースアニメーションと称した．インバースアニメーションは，映像からその生成過程を探る手法である．今後は動作の測定だけでなく動作のシナリオや演技者の意図まで推定できることを目標としたい．

謝辞 映画の使用にご協力いただきましたユニバーサルスタジオに感謝いたします．

参考文献

1) 天谷賢治，原 裕二，青木 繁：逆解析手法による 3 次元人体運動の再構成，機械学会論文集（C 編），Vol.63, No.608, pp.1167–1171 (1997).
2) Barron, C. and Kakadiaris, I.A.: Estimating anthropometry and pose from a single image, *IEEE CVPR00*, pp.669–676 (2000).
3) Bregler, C. and Malik, J.: Tracking people with twists and exponential maps, *CVPR98*, pp.8–15 (1998).
4) Campbell, L.W. and Bobick, A.F.: Recognition of human body using phase space

constraints, *Proc. ICCV95*, pp.624–630 (1995).
5) DeLoura, M.（編），川西裕幸，狩野智英（訳）：*Game programing gems*, ボーンデジタル (2001).
6) Gavrila, D.M. and Davis, L.S.: 3-D model-based tracking of humans in action: A multi-view approach, *IEEE CVPR96*, pp.73–80 (1996).
7) 浜田康志，島田伸敬，白井良明：遷移ネットワークに基づく多視点画像時系列からの手指形状推定，信学論，Vol.J85-D-II, No.8, pp.1291–1299 (2002).
8) Hannan, S.A., Endou, D. and Yamamoto, M.: Director-oriented shot assessment and evaluation in virtual cinematography，映像情報メディア学会誌，Vol.59, No.4, pp.592–603 (2005).
9) Haritaoglu, I., Harwood, D. and Davis, L.S.: W^4: Who? When? Where? What? A real-time system for detecting and tracking people, *FG'98*, pp.222–227 (1998).
10) Howard, S., Mitchell, M., Bridges, H. and Boodman, T.C.（著），大場啓蔵，森田明春，竹村憲一，田邉直美（訳注）：Gone with the wind 風と共に去りぬ（上・下），南雲堂 (1994).
11) Ivanov, Y.A. and Bobick, A.F.: Recognition of visual activities and interactions by stochastic parsing, *IEEE PAMI*, Vol.22, No.8, pp.852–872 (2000).
12) 岩井儀雄，八木康史，谷内田正彦：単眼動画像からの手の 3 次元運動と位置の推定，信学論，Vol.J80-D-II, No.1, pp.44–55 (1997).
13) 亀田能成，美濃導彦，池田克夫：シルエット画像からの関節物体の推定法，信学論，Vol.J79-D-II, No.1, pp.26–35 (1996).
14) Katz, S.D.: *Film directing shot by shot: Visualizing from concept to screen*, Michael Wiese Production/Focal Press (1991).
15) 川島宏彰，松山隆司：連続状態モデル間の相互作用に基づく多視点動作認識，信学論，Vol.J82-D-II, No.12, pp.1801–1812 (2002).
16) Kim, T., Park, S. and Shin, S.: Rhythmic-motion synthesis based on motion-beat analysis, *ACM SIGGRAPH*, pp.392–401 (2003).
17) 近藤拓也，山際貴志，山中光司，山本正信：動画像からの動作感性情報の抽出，信学論，Vol.J80-D-II, No.1, pp.247–255 (1997).
18) Menache, A.: *Understanding motion capture for computer animation and video games*, Morgan Kaufmann (2000).
19) 三富文和，藤原冬樹，山本正信，佐藤泰介：習慣的な行動の確率文脈自由文法に基づくベイズ推定，信学論，Vol.J88-D-II, No.4, pp.716–726 (2005).
20) 大野　宏，山本正信：文字認識手法を用いた固有平面上での動作認識，情報処理学会論文誌，Vol.40, No.8, pp.3134–3142 (1999).
21) 大田佳人，山際貴志，山本正信：キーフレーム拘束を利用した単眼動画像からの人間動作の追跡，信学論，Vol.J81-D-II, No.9, pp.2008–2018 (1998).
22) 大谷　淳，岸野文郎：遺伝的アルゴリズムを用いた多眼画像からの人物の姿勢のモデルベース推定，映像情報メディア，Vol.51, No.12, pp.2107–2115 (1997).
23) Rfros, A.A., Berg, A.C., Mori, G. and Malik, J.: Recognizing action at a distance, *Proc. ICCV03*, pp.726–733 (2003).
24) 塩沢由典：複雑さの帰結，NTT 出版 (1997).
25) 島田伸敬，白井良明，久野義徳：確率に基づく探索と照合を用いた画像からの手指の 3 次元姿勢推定，信学論，Vol.J79-D-II, No.7, pp.1201–1217 (1996).
26) Stone, M., et al.: Speaking with hands: Creating animated conversational characters from recording of human performance, *ACM SIGGRAPH*, pp.506–513 (2004).
27) 千　宗室，裏千家茶道教科　点前編全 17 巻，淡交社 (1976).
28) 高橋勝彦，関　進，小島　浩，岡　隆一：ジェスチャー動画像のスポッティング認識，信学論，Vol.J77-D-II, No.8, pp.1552–1561 (1994).
29) Taylor, C.J.: Reconstruction of articulated objects from point correspondences in a single uncalibrated image, *IEEE CVPR00*, pp.677–684 (2000).
30) Tomasi, C., Patov, S. and Sastry, A.: 3D tracking = Classification + Interpolation, *Proc. ICCV03*, pp.1441–1448 (2003).
31) 和田俊和，佐藤正行，松山隆司：選択的注視に基づく複数対象の動作認識，信学論，Vol.J82-D-II, No.6, pp.1031–1041 (1999).
32) 八木下勝利，山本正信：シーン拘束を用いた人間動作の高精度動画像追跡，映像情報メディア，Vol.52, No.3, pp.331–336 (1998).
33) Yamamoto, M. and Koshikawa, K.: Human motion analysis based on a robot arm model, *IEEE CVPR91*, pp.664–665 (1991).
34) 山本正信，川田　聡，近藤拓也，越川和忠：ロボットモデルに基づく人間動作の 3 次元動画像追跡，信学論，Vol.J79-D-II, No.1, pp.71–83 (1996).
35) 山本正信，星　昌人，下山　功，五十嵐達也：サウンドとモーションの対応付けからのキャラクタの動作生成，信学技報（PRMU）(May 2001).
36) 山本正信，八木下勝利，古山隆志，大久保直人，星　昌人，星野準一，山中　一：映画からの俳優の演技の測定とアニメーションでの再利用，日本バーチャルリアリティ学会論文誌，Vol.7, No.4, pp.503–512 (2002).
37) 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，Vol.J88-D-II, No.7, pp.1153–1165 (2005).
38) 大和淳司，大谷 淳，石井健一郎：隠れマルコフモデルを用いた動画像からの人物の行動認識，信学論，Vol.J76-D-II, No.12, pp.2556–2563 (1993).

(平成 17 年 8 月 31 日受付)
(平成 18 年 3 月 17 日採録)

（担当編集委員 島田 伸敬）

山本 正信（正会員）

昭和 26 年生．昭和 50 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．平成元年〜2 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教授．動画像処理，コンピュータビジョン等の研究に従事．工学博士．昭和 62 年情報処理学会研究賞受賞．電子情報通信学会，映像情報メディア学会，IEEE-CS，ACM 各会員．