2007－CVIM－158 (16)
2007／3／19

3次元モデルを利用した人物モーションキャプチャ技術

～最新LSIを利用したリアルタイム処理の実現～

○谷口恭弘* 岡田隆三* 檜田和浩** 池司* 福田悦生** 近藤伸宏**

*（株）東芝 研究開発センター **（株）東芝 セミコンダクター社

[概要] 近年、コンピュータの高速化とメモリーの大容量化に伴い、認識対象に関する詳細な3次元モデルを用いた認識手法が数多く提案されている。特に、認識対象が人間である場合には、人体の基本的な機構は人種、性別によらず共通であるため、標準的な3次元人体モデルとその変形によって人の動きや姿勢を認識することが可能となる。
本論文では、人物の3次元モデルを用いた顔の詳細なトラッキングと全身のモーションキャプチャの2つのアプリケーションを中心に、人を認識するための最新技術を紹介する。ここでは特に人物のリアルタイム認識に焦点を当て、近年一般化してきているマルチコアプロセッサを利用した高速な人物の認識手法について述べる。

Real-time Human Motion Capture using 3D Human Body Model

○Yasuhiro Taniguchi* , Ryuzo Okada* , Kazuhiro Hiwada** ,
Tsukasa Ike , Etsuo Fukuda** and Nobuhiro Kondo**

* Toshiba R&D Center ** Toshiba Semiconductor Company

Abstract: This paper proposes a method for marker-less human motion capture using 2D or 3D human models. We introduce three real-time applications based on human motion capture. One is a gesture recognition system using AdaBoost algorithm. Another is a virtual make-up system using 3D morphable face model. The other is a marker-less human body motion capture system based on tree-based filtering. Our algorithm utilizes morphable 3D models consisting of a combination of 3D linear bases, and it is extremely well suited to the task of fitting the 3D model to the target object in real time. Experimental results show the effectiveness of these methods based on 3D models.

1. はじめに

画像による人物の検出は、コンピュータインターフェース、ゲーム等のエンターテイメント、画像監視、モーションキャプチャなど様々な応用が考えられ、盛んに研究が行われている。

特に、近年はカメラ等の撮像装置の高機能化とコンピュータの高性能化により、リアルタイムに正確な人物の検出を行うことが可能になってきている。

カメラで撮影された画像から人物を検出する場合、2次元の見え方モデル（参照画像・データ）を元に画像中の人物を検出する方法と3次元の人体モデルを元に検出する方法が考えられる。

2 次元の画像をベースに人の顔を検出する手法としては、AdaBoost 法を用いた手法がよく利用されている。この手法では、モデルを作成するための学習には多くの時間が必要であるが、照合自体は比較的短時間で行うことが可能である。このため、デジタルカメラなどで撮影された人物の顔領域の明度を元に画像全体の明度を自動補正する機能として実装している製品などに利用されている。

これに対して、3 次元のモデルを利用して顔を含む人体の形状や姿勢を検出するためには、3次元モデルの変形と3次元から2次元へのデータの投影などが必要となり、比較的多くの処理時間が必要である。

人体のパーツのうち、顔の形状の変形は腕や足などの形状の変化に比べてその変形量は小さく、目や口の動きと表情の変化が主な変形要素となる。そこで、顔の 3 次元データを扱う場合には、これらの変形が可能な 3 次元のパッチモデルを用いることによって顔の検出が可能となる。また、面長や丸い顔など顔の形状の違いについては、代表的な形状の標準モデルを準備し、ユーザーの顔形状に最も近い標準モデルと各ユーザーの顔のパーツの位置のずれから標準モデルを微調整することによって、各ユーザーに合った 3 次元顔モデルを生成することが可能となる。

人間の全身は多数の関節からなる高い自由度を持った関節物体であり、インターフェースやゲーム等実時間性が要求されるアプリケーションでは、高次元の姿勢状態空間を効率よく探索して最適な姿勢を求めることが重要な課題となる。

フレーム間の姿勢変化を検出し、各フレームでの人物の姿勢を推定する手法は、前フレームの姿勢の近傍のみを処理することにより計算量を低減することができるが、手動で正確な初期姿勢やキーフレームの姿勢を与えなければならないといった問題や、推定誤差が蓄積するという問題がある。このようなフレーム対応付け問題に対して、姿勢の状態空間を表現する木構造を利用し、ベイズ推定の枠組みに基づいて、効率的かつ安定に姿勢を探索する手法(Tree-based filtering)[1]が提案されている。

本論文では、画像特徴の類似度に基づいて構成した木構造に運動モデルを導入し、効率的で安定な姿勢推定を行う。しかし、カメラの数が少ない場合には、姿勢推定を行う際に部位の隠蔽が問題となる。そこで提案手法では、部位の隠蔽を考慮して予測モデルを切り替えることによって、隠蔽に対して安定な姿勢推定を行う。また、カメラの設置の容易さやシステム規模を小さくすることを念頭において、固定の単眼カメラを使用する。さらに、背景や照明条件はある程度制御できるという前提の下、背景差分によるシルエットを用いる。本論文では 2 つのシルエット間の類似度の評価値として、シルエットの中心に近いほど重みを大きく設定した排他論理和を、部位の太さを考慮して正規化した、正規化中心重み付排他論理和を使用する。この評価値により、腕など細い部位と胴体など太い部位を等価に扱うことができ、細い部位の推定の安定性が向上する。

以下、2 章では 2 次元モデルを利用した人物検出の例として近年よく利用されている AdaBoost 法をジェスチャ認識に利用したシステムについて述べる。次に、3 章では顔の 3 次元モデルを利用した仮想化粧システムについて述べ、4 章では Tree-based filtering を用いた実時間で動作する人物モーションキャプチャシステムについて述べる。

2. 2 次元モデルを用いたジェスチャの検出[2]

ジェスチャによって電子機器を制御する場合、高速かつ安定的に手の形状や動きを検出することが重要となる。画像中から手の領域を検出する手法は数多く提案されているが、ここでは高速に動作させるために並列度の高いプログラムを作成可能な AdaBoost[3]を用いた手法について述べる。本手法では、手を検出する際に効果の高い検出器の形状を多数の手画像を用いて学習し、学習によって得られた検出器を用いて画面中から手を検出する。

AdaBoost 法は当初形状変化の少ない人間の顔を検出する手法として提案されたが、ジェスチャ認識においても腕の動きではなく手の形に注目することによって本手法を適用することが可能となる。

実際の利用場面では、画像中の手の大きさはシーンによって大きく異なるため、学習した検出器のサイズを複数変更することによって対応した。一般に、形状やサイズは互いに独立した特長と考えられるので、マルチコアプロセッサで実行する場合に異なった演算器で並列して処理することが可能となる。

図 1. 認識対象とする手姿勢

今回試作したシステムでは、図 1 に示した 3 種類の手形状および人の顔を検出することが可能である。3 種類の手形状を用いることによって、複数の機能を実現することが可能となる。本システムでは、これらの 3 種類の手姿勢について、それぞれ AdaBoost によって生成した識別器を用いてビデオカメラから取得した画像を走査することにより手を検出する。また、識別器を拡大縮小することにより 12 種類の大きさの識別器を生成し、様々な大きさの手を検出する。識別器を構成する特徴としては、顔検出に用いられる Joint Haar-like 特徴[4]を用いている。これは、Viola-Jones の顔検出手法で用いられる Haar-like 特徴 (2 個の矩形領域内の画素値平均の差) を拡張したもので、複数の Haar-like 特徴間の相関に着目して検出を行うことで、検出性能を向上させている。これにより、輪郭等の手を特徴づける部分に重点を置いて少ない計算量で検出処理を行えるため、高速かつ環境変化に頑健な検出を実現できる。

今回の試作システムは、多くの人物が存在するシーンに対しても安定に動くように、全画面の中から手を検出しているが、一般的な用途を考えた場合には、手の検出位置や大きさなどはある程度限定することが可能である。この場合は、より少ない数のプロセッサで同様の処理を実現することが可能となり、顔認識などの他のアルゴリズムと協調したインテリジェントなシステムの構築が可能になると考えられる。試作したジェスチャ認識システムを図 2 に示す。

図 2．ジェスチャによる AV 機器操作の例 ジェスチャ認識により、選択、再生、停止などを行うことができる。

3. 3 次元顔モデルを用いた顔の検出

画像中から人物の顔や手領域を検出することが目的である場合は、前章で述べたように 2 次元のモデルを元に領域を検出することで対応可能であるが、検出した顔領域に対して化粧や髪型のシミュレーションなどを施す場合には顔領域の 3 次元情報を獲得する必要がある。

3.1 3 次元モデルを利用した顔領域の検出

本システムでは、まず、利用者の顔をカメラで撮影し、その映像を元に利用者の顔に最も近い顔モデルを選択する。

顔モデルの選択時は、入力画像から自動的に

検出された目と口の領域情報と操作者が手入力するあごの輪郭情報を元に複数の顔モデルから最も利用者の顔に近いモデルを選択する。顔モデルは大きさが正規化されたデータとして保存されており、細い顔、丸い顔、エラの張った顔など、人の顔の特徴を表す 11 種類のモデルを用意した。

利用者の顔に最も近い顔モデルが選択されると、この顔モデルと入力画像を元に顔トラッキグや化粧・髪型のシミュレーションを行う。ここで、顔トラッキング処理では、顔の位置、向き、表情を認識し、その結果は 3 次元仮想化粧や髪型シミュレーションに利用される。仮想化粧や髪型シミュレーションを行う場合は、あらかじめ用意した化粧の画像や髪型の画像と、カメラで撮影した顔の画像を自然に見えるように合成し、利用者の前に設置されたモニタに出力する。仮想化粧、髪型シミュレーションの処理はそれぞれ独立に行うことが可能であり、マルチコアプロセッサを用いることによって、化粧と髪型両方のシミュレーションを同時に行うことができる。

3.2 顔トラッキング

顔トラッキング処理では、最初に、目、鼻、口を自動認識し、顔の位置を把握する。次に、テンプレートマッチングによって、顔全体から抽出した、およそ 500 個の点の動きを追跡する。この点の動きから顔の向き、表情を認識することができる。

テンプレートマッチング処理では、ある点の周囲 8x8 画素の画像が、周囲 20x20 の領域のどこに移動したかを判定しており、8x8 の画像同士のフィルタリング処理が行われる。したがって、1 回のテンプレートマッチングでは、169 回のフィルタリング処理が行われることになる。さらに、この処理は、追跡を行う 500 個の点全てに対して行うため 1 フレームの処理では、84,500 回のフィルタリング処理を行う必要がある。

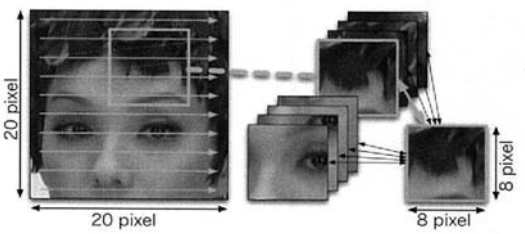


図 3．テンプレートマッチング処理

シングルコアのプロセッサでは、全ての処理を一つのコアで順次実行しなければならない。ところが、この処理は 500 個の点について、どの順序で処理しても良く、並列に処理することが可能である。このような処理では、マルチコアプロセッサを用いると複数の処理を同時に行うことができる。さらに、この処理では SIMD 命令を使用することにより、並列処理を行わない場合に比べると、10 倍以上の処理性能を達成することができる。

3.3 3 次元仮想化粧

3 次元仮想化粧は、テクスチャマッピング技術を用いて実現する。このシステムでは、利用者の顔の表面形状をおよそ 6000 個のポリゴンと呼ばれる三角形で表現しており、その表面形状に化粧の色をのせることで、あたかも化粧をしたかのような画像を作り出している。

一般に、テクスチャマッピングは、市販のコンピュータではグラフィックプロセッサと呼ばれる専用のＬＳＩを利用して処理していることが多いが、本システムでは、プロセッサ内の演算器を利用してソフトウェアで実現した。

3.4 髪型シミュレーション

髪型シミュレーションは、イメージベーストレンダリング処理を用いて実現している。この手法はコンピュータグラフィックスの主要な処理であり、本アプリケーションでは、入力画像と過去に撮影済みの髪型画像という実写同士の融合に利用した。

髪型シミュレーションでは、変更したい髪型の画像データベースを用いる。このデータベー

スには、複数の髪型それぞれに対して複数の方向を向いた画像が登録されている。しかし、利用者が現在向いている方向とまったく同じ方向を向いている画像が登録されているとは限らないので、入力画像の顔の向きに近い複数の画像を画像データベースから選択し、現在の顔の向きに合った髪型の画像を生成している。このような画像を用いた大容量データを必要とする処理には、高速内部バスや、高速メモリインターフェースが有効であり、高速なマルチコアプロセッサを用いる場合でもデータ転送が演算処理の妨げにならないように制御することが必要である。

3.5 実験結果

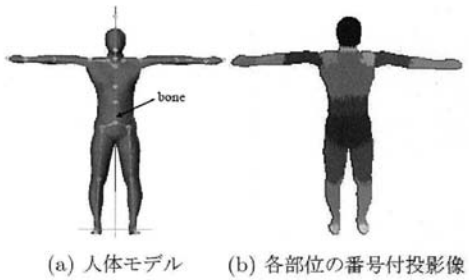
試作した仮想化粧システムによる実験結果を図4に示す。本システムによって、擬似的な化粧、髪型変更体験が実現されていることが示されている。

図4. 処理結果 擬似的な化粧、髪形変更体験をしている様子。

このアプリケーションでは、3.2〜3.4で述べた処理を、1秒間に30フレーム処理しなければならない。しかも、ユーザーの顔や表情に追従するインタラクティブなアプリケーションのため、リアルタイムに処理する必要がある。さらに、市販のコンピュータでは、コンピュータグラフィックス用の専用プロセッサが実行する処理まで含まれている。我々は、これらの処理が8つのコアを持つマルチコアプロセッサによって処理可能なことを実証した。

4. 3次元人体モデルを用いた姿勢推定[5]

人体は、顔のような微妙な形状変化だけでなく、手足などの関節構造による大きな形状変化を伴って稼働するため、顔に比べて非常に大きなモデル空間の中から最適な姿勢を推定する必要がある。本システムでは、洋服のJIS規格(JISL4004およびJIS L4005)に基づいて、図5に示すような男性10体、女性14体の典型的な体型の標準人体モデルを使用し、それぞれについて姿勢の木構造やシルエットを計算して姿勢辞書を作成した。また、追跡時には、追跡対象人物の身長と胸囲を用いて、最も体型の近い標準体型モデルを選択して利用した。



(a) 人体モデル　(b) 各部位の番号付投影像

図5. 人体モデルの例

4.1 画像特徴に基づく姿勢の木構造の作成

認識を行う姿勢の状態空間は非常に広いため、市販のモーションキャプチャシステムを用いて姿勢のサンプルを多数収集し、これら姿勢サンプルの集合を姿勢の状態空間Rと定義する。

木構造の階層$l-1$において、状態空間がN_{l-1}個の重なりのない部分$\{S^{i(l-1)}\}_{i=1}^{N_{l-1}}$に分割されているとする。つまり、$R=\cup_{i=1}^{N_{l-1}} S^{i(l-1)}$である。$S^{i(l-1)}$を、下位の階層$l$において以下のように分割する。ただし、最上位階層$l-1=0$においてはノード数$N_o=1$で、その下の階層

$l=1$では状態空間全体が分割対象の空間となる($S^{10}=R$)。

実際に木構造を生成するためには、以下の処理をあらかじめ定められた階層に達するまで再帰的に実行する。

(1) $S^{i(l-1)}$に含まれる全ての姿勢サンプルの平均値に最も近い姿勢サンプルを、木構造のノードを代表する姿勢(代表姿勢) として、新しいノードを生成する。
(2) $S^{i(l-1)}$ に含まれる姿勢の中で、現在までに選択されている全ての代表姿勢からの画像特徴距離(シルエット間の距離) が閾値T_lより大きく、代表姿勢から平均距離が最も大きい姿勢を選択する。このような姿勢が存在すれば、新たなノードを生成してその代表姿勢とし、2 へ戻る。存在しなければ、3 へ進んで代表姿勢の選択を終了する。
(3) $S^{i(l-1)}$の中で、代表姿勢に選ばれていない姿勢について、最も画像特徴距離が小さい代表姿勢を選択し、代表姿勢の属するノードに登録する。各ノードに属する姿勢サンプルの集合が分割された状態空間S^{jl}となる。
(4) 各ノードに属する姿勢サンプルの平均値に最も近い姿勢サンプルを、代表姿勢として再定義する。

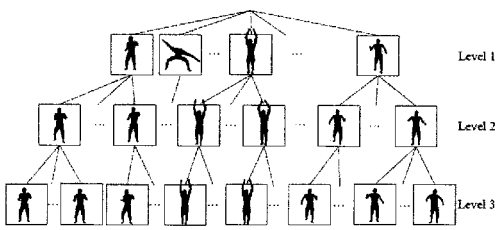


図 6．姿勢の木構造モデルの例

作成された木構造の例を図 6 に示す。

次に、姿勢サンプルから木構造を作る際に用いる画像特徴であるシルエットを抽出する方法について述べる。

様々な姿勢のシルエットを生成するために、あらかじめ姿勢推定を行う人物の体の 3 次元形状モデルを次のように取得しておく。

(1) 追跡対象人物の表面形状をポリゴンで近似する。画像への投影にかかる計算量の削減のためのポリゴン数は数千程度に削減する。
(2) 姿勢を変化させるための骨格構造を埋め込む。
(3) 表面形状を変形させるため骨格構造の各部位にポリゴンを分類する。

図 5 に示したような人体モデルを、木構造の各ノードの代表姿勢に変形させ、全ポリゴンの隠蔽を考慮しながら画像に投影する。このとき、各ポリゴンが属している部位の一意に定められた番号を画素値とすることにより、各部位の番号付投影像を得ることができる。シルエットはこの投影像を背景とそれ以外で 2 値化すれば得られる。

4.2 隠蔽を考慮した運動モデル

4.1 節で述べたように、人体の 3 次元モデルの画像への投影によって部位の番号付投影像が得られる。この投影像を用いて、各部位の面積(画素数) が得られる。ある姿勢xを持つ人体モデルの部位jの投影面積をA_jとすると、投影面積が閾値T_Aより小さい部位を隠蔽部位とする。隠蔽情報は、以下のようになる。

$$F_{occ}(x)=\{f_j\}_{j=1}^{N_b},\quad f_j=\begin{cases}1 & if\ A_j<T_A\\ 0 & else\end{cases}$$

ここで、N_bは人体の部位の数である。

運動モデルは、過去のフレームの姿勢から現在のフレームの姿勢への遷移確率分布で与えられ、一次マルコフモデルを仮定して$p(x_t \mid x_{t-1})$となる。本論文では、認識する動きに対する制約を少なくするため、運動モデル

としてランダムウォークモデルを採用する。すなわち、前フレームの姿勢x_{t-1}を平均値とする正規分布を運動モデルとして用いる。

$$p(x_t \mid x_{t-1}) \sim N(x_{t-1}, \Sigma)$$

ここで、Σは各部位の運動の速さを考慮して実験的に定めた。

提案手法では、正規分布の分散を部位の隠蔽情報$F_{occ}(x)$を用いて、隠蔽が起こっている部位に対しては、分散の大きい運動モデルを選択する。

$$\Sigma = diag(\sigma_j'^2) \quad , \quad \sigma_j' = \begin{cases} \sigma_j & if\ f_j = 0 \\ m\sigma_j & if\ f_j = 1 \end{cases}$$

ここで、σ_jは、各関節に関する運動モデルの標準偏差、m>1は隠蔽が起こっている場合の標準偏差の拡大率を表すパラメータである。本論文の実験では、m=5を用いている。これにより、隠蔽部位についてのみ運動モデルに合致しない姿勢変化を許容する。

4.3 画像特徴に基づく Tree-based filtering

画像特徴に基づく Tree-based filtering においても、4.1 節で生成した木構造を用いて、最下層の各ノードの事後確率を従来の手法[1]と同様に計算することができる。

提案手法では、木構造の最下層において最も高い解像度で状態空間の分割が行われていても、状態空間の分割を画像特徴距離に基づいて行っているため、腕が胴体に隠蔽されて見えていない姿勢等、大きく異なる姿勢が同じノードに登録されている場合がある。

そこで、最下層のノードに関する事後確率が閾値より大きいノードについて、ノードに含まれる各姿勢サンプルの事後確率を計算し、事後確率が最も高い姿勢サンプルを現在のフレームの推定結果とする。姿勢の木構造は、画像特徴距離に基づいて構成されているため、最下層の同じノード内の姿勢サンプルの尤度は一定とみなし、ノードに関する尤度を用いた。

最も解像度の高い階層においても一意に定まらない姿勢は、画像特徴からは判断できないため、運動モデルによる姿勢の時間的連続性に基づいて現在の姿勢を定めることは妥当である。つまり、最下層の同一ノード内では、運動モデルにより事前確率分布のみ変化して、各姿勢サンプルについての事後確率分布が計算される。ただし、初期確率分布は既知とする。初期確率分布に一様分布を仮定するなど、一意な初期姿勢が与えられていない場合、初期フレームにおいて画像による観測から姿勢を一意に決めることができなければ、あいまい性が残ったままとなる。本論文では、初期姿勢は正面向きの直立姿勢と仮定し、そのような姿勢を平均値とする正規分布を初期確率分布とする。

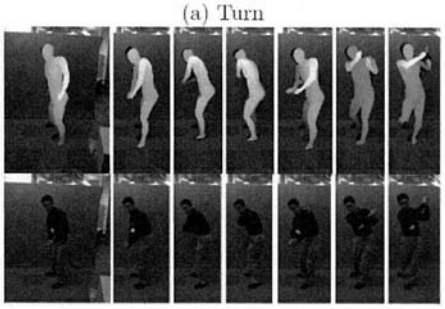
4.4 実験結果

図 7 は、様々な動作を含む画像列に対して追跡を行った結果から、2 種類の動作に関する結果を切り出して表示したものである。この実験に用いた画像列の長さは 2 分間でフレームレートは 15 fps である。姿勢サンプル数は 57、136 姿勢で、シルエット距離に基づいて木構造を生成した結果、図 6 と同様に生成した木構造モデルにおいて、ノード数は上位層から順に、7,828、26,805、44,108 ノードとなった。

図 7(a) は、カメラの光軸に垂直な軸周りの回転運動で、腕が隠蔽されたり、正面と背面の区別が難しい例であるが、正しく姿勢推定が行われている。図 7(b) は、ゴルフスイングをカメラに向かって行っている。カメラから遠い方の右腕が隠蔽されているが、正しく推定できた。

提案手法を用いて、図 8(a) のようなオンライン姿勢推定システムを構成した。カメラで取得した画像を、OpteronTM 280 を 2 基搭載したハイエンド PC、またはマルチコアプロセッサを用いたシステム(図 8(c) 参照) に入力して姿勢推定を行った。その結果、ハイエンド PC

(a) Turn



(b) Golf swing

図 7．実験結果

では、1 フレーム平均 128 ms 程度の処理時間を達成した。一方、本実験で用いた Cell は、3.2GHz で動作し 8 個のコアを搭載したマルチコアプロセッサであり、1 フレームあたりの平均処理時間は、86 ms を達成した。

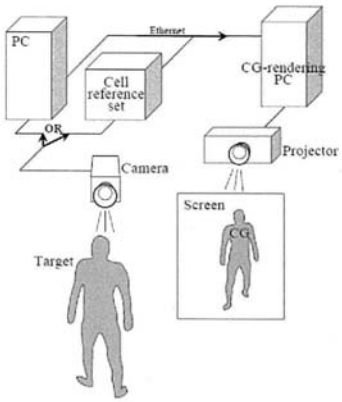
5. おわりに

近年のマルチコアプロセッサの発達により、これまではリアルタイムで扱うことが困難であった認識対象に関する詳細な 3 次元モデルを利用することが可能になってきた。

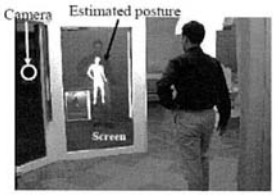
本論文では、2 次元モデルを用いた人物のジェスチャ認識手法について述べた後、3 次元モデルを用いた人物の検出手法について述べた。

3 次元モデルを用いたアプリケーションとしては、顔の 3 次元モデルを利用した仮想化粧と全身のモーションキャプチャの 2 つのシステムを紹介した。

これらのシステムは、人物の高速な認識に焦点を当て、近年一般化してきているマルチコプロセッサを利用することによりリアルタイムでの処理を可能としている。



(a) System organization



(b) System overview

(c) Cell reference set

図 8．オンライン姿勢推定システム

参考文献

[1] B. Stenger, A. Thayananthan, P. H. S. Torr and R. Cipolla: "Filtering using a tree-based estimator", Proc. of ICCV, Vol. 2, pp. 1063–1070 (2003).

[2] 池司等：マルチコアプロセッサを用いたリアルタイムハンドジェスチャ UI,画像の認識・理解シンポジウム」 2006 年 7 月

[3] P. Viola, M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," Proc. of CVPR, pp.511-518, 2001.

[4] T. Mita, T. Kaneko, O. Hori, "Joint Haar-like Features for Face Detection," Proc. Int. Conf. on Computer Vision, pp.1619-1626, 2005.

[5] 岡田等：シルエットを用いた Tree-Based Filtering による人体の姿勢推定,画像の認識・理解シンポジウム」 2006 年 7 月