ヒューマンインタフェース　95－13
コンピュータビジョンと　129－13
イメージメディア
（２００１．９．１３）

View and Motion-based Aspect Modelによる人間行動認識

柏　良輔†　　西村　拓一‡　　石黒　浩†
和歌山大学システム工学部情報通信システム学科†
新情報処理開発機構（RWCP）‡

人間のジェスチャーを本質的に理解するための新しいモデリング手法を提案する.そのモデルを View and Motion-based Aspect Model（VAMBAM）と呼称する．これは全方位画像による「動きの見え方」に基づいたモデルであり，環境に配置された複数の全方位視覚センサからなる全方位視覚システム（DOVS）の中で，人物の位置および回転に依存せずにジェスチャーの認識を実現可能とするものである．DOVS は実世界の監視や認識をするための知覚情報基盤の原型であり，本論文ではVAMBAM の概念に加えて，構築したモデルが DOVS 環境下においてどのようにロバストで実時間性を持った認識が行えるかを示す.

View and Motion-based Aspect Models for Recognition of Human Motion Behaviors

Ryosuke KASHIWA †　Takuichi NISHIMURA ‡　Hiroshi ISHIGURO †
Department of Computer and Communication Sciences, Faculty of Systems Engineering, Wakayama University †
Real World Computing Partnership (RWCP) ‡

This paper proposes a new model for gesture recognition. The model, called View and Motion –based Aspect Models (VAMBAM), is an omnidirectional view-based aspect model based on motion-based segmentation. This model realizes location-free and rotation-free gesture recognition with a distributed omnidirectional vision system (DOVS). The distributed vision system consisting of multiple omnidirectional cameras is a prototype of a perceptual information infrastructure for monitoring and recognizing the real world. In addition to the concept of VAMBAM, this paper shows how the model realizes robust and real-time visual recognition of the DOVS.

１．はじめに

物体認識を行う手法として，３次元モデルと View-based Model を用いる方法がある．３次元モデルを用いる方法は，物体の形状や線画が得られることを対象にしており，それは制限された環境下では有効であるが，より一般的な開かれた現実世界のもとではうまく働かない[9]. これに対して, View-based Model は観察された画像群を用いた見え方の変化モデルであり，現実世界を対象としている．しかしこれらの手法はいずれも静的な構造抽出にとどまっている.

我々は，物体の持つ意味は「構造」ではなく，その物体の動きや，動的変化，他の物体との相互作用による「動きの見え方」により定義できるものだと考える．つまりは対象となる物体を連続した時間で観察することによって，本来のアスペクトモデルを構築できるものと考える．本論文ではその基本となるアイデアと開かれた環境下での行動認識についての提案手法を述べる.

本システムでは人間を観察するセンサとして，全方位視覚センサを複数台用いている．図１に実際に使用した全長６[㎝]程度の小型全方位視覚センサを示す

[1][8]．このセンサは実時間で 360 度の画像が取得できるため，広い環境に複数台配置することで環境内全ての画像を隙間なく取得することができる．そのイメージを図２に示す．我々はこのシステムを Distribute Omnidirectional Vision System（DOVS）と呼び，この DOVS を基盤として，コンピュータと多数のセンサ間をネットワークで結ぶ「知覚情報基盤」の開発を最終目標としている．

図１：全方位視覚センサ

この DOVS が配置された環境内で，人間の動きを実時間で追跡しその行動を認識することを考えと，当然，人間は環境内を縦横無尽に動きまわるため，人間の立っている位置と向いている方向に依存しない認識を考える必要がある[10]．

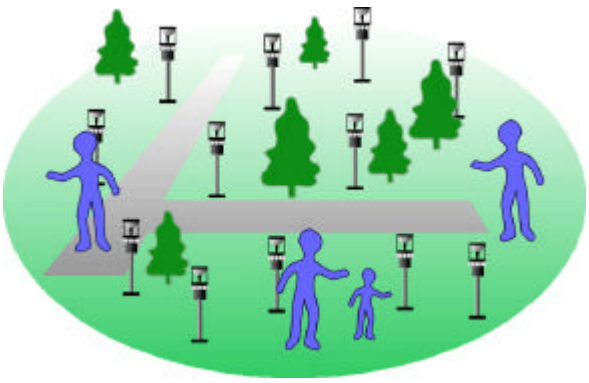

図２：Distribute Omnidirectional Vision System（DOVS）

本論文では，以上に述べた「動きの見え方」による認識と，人物の位置や方向に依存しない新しい行動認識のモデルとして，View and Motion-based Aspect Model（VAMBAM）を提案し，その有効性の検討のため，人間のジェスチャー認識システムを構築した．

以降，２節にて VAMBAM の詳細を述べ，３節にて構築したジェスチャー認識システムの解説をする．さらに４節にて評価実験を行い，本手法の有効性を検証し，５節にてまとめとする．

２．View and Motion-based Aspect Model

本システムにおいて，ジェスチャー認識の行程は２つに分けることができる．ひとつはモデルを構築するモデリング部分．もうひとつは，構築したモデルと対象の動作をマッチングする認識部分である．

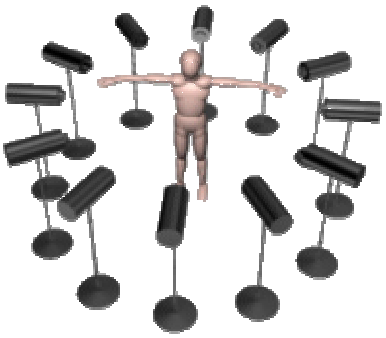

図３：カメラ 16 台によるモデリング

図４：DOVS 環境でのジェスチャー認識

モデリング部分において，VAMBAM は図３に示すように人間の周囲に設置された 16 台のカメラの中で行う数種類のジェスチャーを，システム内に蓄えることで構築される．

そして認識部分では，図 4 のような DOVS 環境に配置した複数台の全方位視覚センサが人間を追跡し，蓄えられた数種類の VAMBAM とのマッチングを行い，ジェスチャーを認識する．

理想的なモデルというのはもちろん全ての方位から

観察したモデルなのであるが，そうするには無限とも言える多数のカメラをオブジェクトの周囲に配置せねばならず，それは不可能である．そのためにVAMBAMはそれを近似したモデルとなる．近似に関して問題となるのはカメラの台数であるが，それは物体の複雑さに依存する．現在のDOVSの目的は人間の行動認識であり，その準備段階の実験として本システムではカメラの台数は16台に固定してある．

16台のカメラを用いたシステムは16個の画像シーケンスと特徴ベクトルを抽出する．ここで用いている16台のカメラにおいて重要度等は設定していない．ジェスチャー認識においては，おそらく前方からの視点の方が横方向からのものよりは重要度が高い等の事例が考えられるため，今後，それぞれの特徴ベクトルに重み付けを行うことで，動きに基づき，かつ最適な方向を選択した認識が可能となることが考えられる．

3．ジェスチャー認識システム

図5を用いて，構築したジェスチャー認識システムの概要を説明する．まずあらかじめ人物の動作を全方向から観察し，この画像系列から特徴ベクトル列を求め，これをVAMBAMとして作成しておく．

次にN個の全方位視覚センサから得られる全方位画像の透視変換を行い，背景差分処理[2][3]により背景を排除し，N眼ステレオ視を用いて人物位置を推定する．さらに，得られた人物位置を手がかりに，特徴モデルと同様な特徴を抽出し，これを認識系への入力とする．この入力はN個のセンサから得られた特徴ベクトル以外に，カメラと人物の位置関係をも含む．

この入力モデルとのマッチングを行うために，時空間連続DP（STCDP）を用いてモデル全体との累積距離を求める．このSTCDPにより，モデルに対する人物の動作について，時間方向に1/2～2倍の伸縮を許し，かつ，動作中にセンサに対する人物の方向が変化しても認識可能となる．

3．1　実時間人間追跡システム

ここで言う人間の位置や角度に依存しない認識を行うためには2つの技術が必要となる．そのひとつが実時間人間追跡システムである．DOVSからの画像情報を用いて，我々は複数のカメラを用いた統計的背景差分処理によるロバストな実時間ステレオシステムを構築した．このステレオ視のことをN眼ステレオ視とする．これは3眼ステレオの拡張である[4]．N眼ステレオ視の実行手順は以下のようになる[5]．

(1) 2眼ステレオ視を行い，任意の2つのセンサによって検出された物体の方位角を表す直線の交点に物体をあらわす円を描き，それを2眼視によって推定された物体の位置とする(図6の3つの黒い円)．ここでは検出する物体は人間であり，円で表されると仮定する．

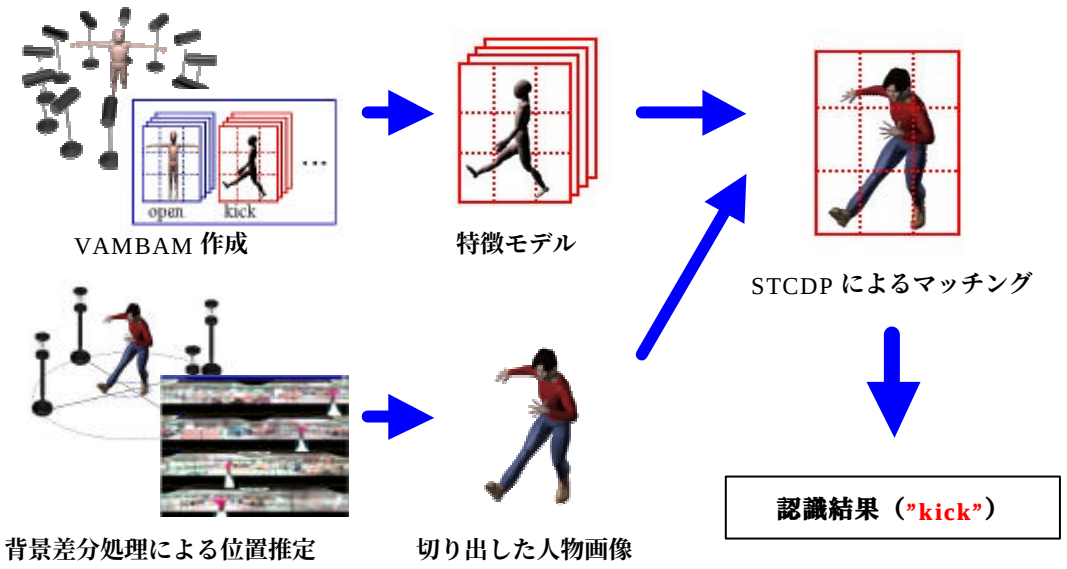

図5：ジェスチャー認識システム概要

（2）N眼ステレオ視においては，N番目のセンサで物体が検出されているかどうかは，それらの円が重なっているかどうかで判定する．重なっている場合には，それらの重心に新たに円を描き，その円をN眼ステレオ視によって検出された物体の位置とする（図６の破線の円）．

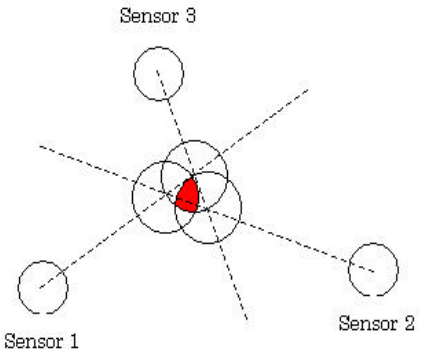

図６：N眼ステレオ視による位置推定

３．２　時空間連続 DP

方向や角度に依存しない認識を行うためのもうひとつの技術が，ここで述べる時空間連続 DP（STCDP）である．STCDP は連続 DP のパスに空間方向の自由度を付加したものである[6]．つまりは，時系列データであるモデルに空間方向の次元を加えたものである．本研究では人物の方向 i_q（１次元）を空間方向の軸とした場合について，図７を用いて説明する．従来の連続 DP は図７における破線のように，モデルのフレーム軸τと入力フレームｔのみであり，従って，用いられる局所パスは３個のみであった．一方，人物の方向変化を許容するためには，図７の一点鎖線のように i_q 軸の変化を許容するパスと時間方向の変化を許容するパスを融合し，９個の局所パスを採用することとする．これによって人物が動作を行う場合，モデルと比べて時間方向に 1/2～２倍の伸縮を許容するだけでなく，人物の方向についても入力１フレームあたり最大２π / N_q の角度変化を許容することになる（図７の実線部分）．さらに大きな角度変化を許容したい場合は，同様に時間方向にシフトした点からパスを貼ればよい．

３．３　特徴抽出領域の決定

本節では，図８に示す特徴抽出を行う領域 A_f を決

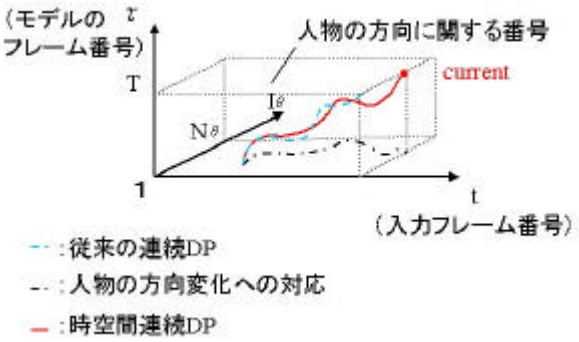

図７：時空間連続 DP（STCDP）

定する手法を説明する．まず，前節の位置推定結果から人物の位置において，横 w[m]，縦 h[m]の映像を切り出し，これを横 N_w[pixel]，横 N_h[pixel]の人物正規化画像に変換する．この処理によって，センサと人物の距離によらず，人物を一定の大きさで捕らえることができる．ここで，人物サイズ正規化画像中の特徴抽出領域は以下のように決定した．まず，背景差分処理により人物領域が１，その他が０となる２値画像を求め，次に，図８のように人物領域に対して垂直方向への射影を行い，閾値 thr_w 以上についての重心点 c_w を特徴抽出領域の水平方向中心位置とする．さらに，水平方向への射影を行い，閾値 thr_h，thr_f から頭頂と足元の垂直方向の位置 h_h，h_f を決定する．この h_f を特徴抽出領域の垂直方向の底部位置とする．

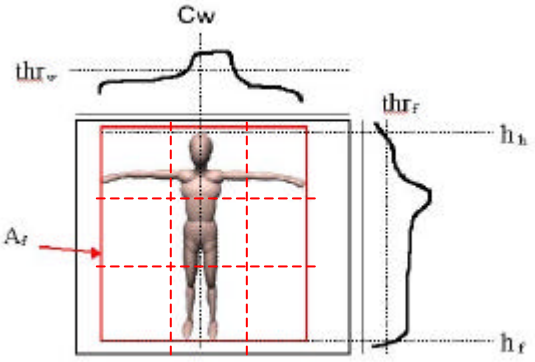

図８：特徴抽出領域の決定

３．４　低解像度特徴の抽出

本研究では画像中の背景差分処理後の人物領域（図８の A_f）を３×３に分割し，各部分領域中の，ある閾値よりも大きい値の pixel の割合を求めた低次元の

特徴を，低解像度特徴として抽出する[7]．この特徴は動きが大きければ，若干の人物位置や動作軌跡の変化および衣服や背景の変化によって，入力画像がモデル作成時と異なったときでもロバストな認識が可能という特徴がある．

４．　実験

４．１　VAMBAM 評価実験

VAMBAM のモデル作成のために，図９に示すように人物を中心とする半径 2[m]の円上に 22.5 度間隔で全方位視覚センサを 16 台配置した．なお全方位視覚センサの高さは 1.5[m]に固定している．なお，ここでは認識のモデルとして図 10 に示す 10 通りの動作を用いた．

図９：実験環境

以下の図 10 における（1）～（10）の各動作を 10 回繰り返して入力画像列を作成した．この実験では，正答数を入力動作総数（10×10）で割った値を認識率とした．ただし，入力動作と同じ認識結果のみを４フレーム（0.4 秒）以上連続して出力した場合を正答とした．ここで，低解像度特徴の分割数 $N_h \times N_w$ は３×３として認識率を求めた．さらに，マッチング時の全方位視覚センサは図９に示すうちの 90 度間隔をなす４台を用いた．この４台の位置はあらかじめキャリブレーションを行っているものとする．

実験は，まず始めに人物が中央で静止しているときの動作を行い，次に VAMBAM の有効性検証のために中央で回転しながらの動作，STCDP の有効性検証のために移動しながらの動作，最後にそれらを組み合わせた移動，回転しながらの動作を行った．入力およびモデル作成時の各全方位画像のサイズは 240×320（4 個の画像で 640×480），フレームレートは 10[frame/s]

図 10：モデル動作

であった．

複数のセンサ入力を用いる効果を調べるため，認識部への入力は，第一のセンサから得られる特徴ベクトルのみの場合と，４台のセンサ全てを用いる場合とで比較した．モデルとのマッチングは一方向のモデルの場合は従来の連続 DP を用い，VAMBAM の場合は連続 DP と STCDP とで実験を行った．表１および表２にその実験結果を示す．

人物が静止しているときは，一方向から捉えたモデルを用いても比較的高い認識率が得られた．これは，モデル作成で用いたセンサと入力時のセンサを一致させたためである．VAMBAM を用いることで認識率が上昇しているが，特に４個の入力情報全てを用いることで認識率が 100%に上昇している．これは前方からの情報だけでは判別困難な「kick」,「throw」,「take」,「bow」の４つの認識率が上昇したためである．従来の連続ＤＰに対する STCDP の効果は，入力が一方向の場合に現れている．これは，同じ動作でも複数回行う中で若干の方向変動が生じたためと考えられ，この変動を STCDP が効果的に吸収したためと考えられる．

人物が移動しているときは，一方向のモデルの認識率が大きく低下している．これは人物の移動に伴って，方向が様々に変化するためである．一方，VAMBAM と連続 DP により認識率は向上するが，４方向の入力を用いても 64％となっている．これに対して，VAMBAM と STCDP を用いることで，人物の方向変動を吸収し，87%の認識率が得られた．これによって，本システムの有効性が示された．

また，本実験で作成した実時間認識システムの各ウ

インドウを図 11，12，13 に示す．図 11 は 4 個の入力された全方位画像および，この 4 個の入力を透視変換し，背景差分処理により人物領域を示したものが図である．図 12 は 4 個の入力情報をもとに認識を行っているウインドウで，図 13 は人物のセンサに対する位置と，人物の最も近くに存在するセンサからの映像および認識結果を表示したウインドウである．

表 1 ：人物静止時の認識結果（%）

	1model，CDP	VAMBAM，CDP	VAMBAM，STCDP
1 input	60	65	83
4 input	-	100	100

表 2：人物移動時の認識結果（%）

	1model，CDP	VAMBAM，CDP	VAMBAM，STCDP
1 input	22	31	62
4 input	-	64	87

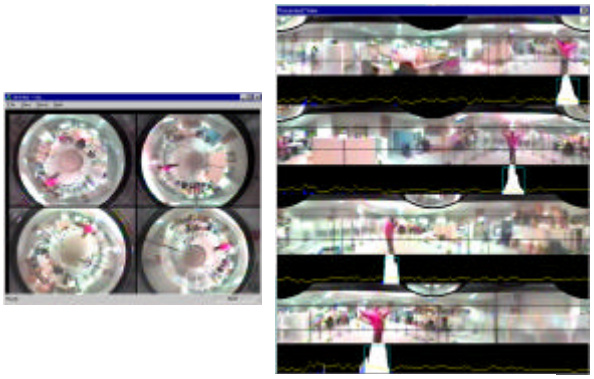

図 11：４つの全方位画像と透視変換画像

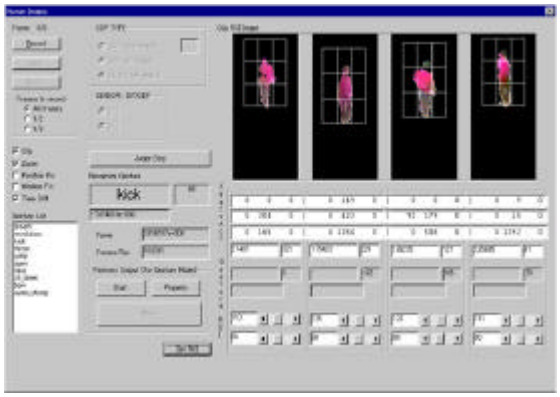

図 12：ジェスチャー認識ウインドウ

図 13：人物の位置と映像および認識結果

４．２　一般環境への適応

前述した実験までは人間のみの行動を扱ったものである．次に，より一般的な環境に適応させるために図 14 に示すように環境内に椅子と机を置いた状態での行動認識，つまりは人間と他の物体との相互作用を認識する実験を行った．

図 14：一般環境への適応

この実験に関して問題となるのは，椅子や机等は環境に固定されてはいるものの，視覚センサの視界が遮られてしまうため，背景差分処理において人間の下半身部分が投影されず，特徴抽出領域が不定となることである．

その問題解決に，ここでは単純に座っている状態では常に頭部が見えており，直立状態のときは常に足元が見えているものとし，図 8 における頭頂の位置 h_h と足元の垂直方向の位置 h_f を基準に特徴抽出領域を決定することとした．まず，図 15 に示すように上段では足元を基準に領域を設定し，直立状態（Standing）時の枠を，下段では頭頂部を基準に領域を設定し，座っている状態（Sitting）時の枠を決定した．次に，そこ

から上下を比較してモデルとの適合度の高い方を選択し，認識を行うこととした．この実験に際しては試作段階ゆえに，歩いている状態からイスに座る「sit-down」，イスから立ち上がる「stand-up」，座っている状態で伸びをする「stretch」の３種類の動作モデルのマッチングではあったが，高い確率で認識を行う事ができた．これは，たとえいくつかの方向のカメラから人間の姿形が確認できていない状況であったとしても，全方向の視点から作成したモデルがあるために認識を行うことができたといえる．

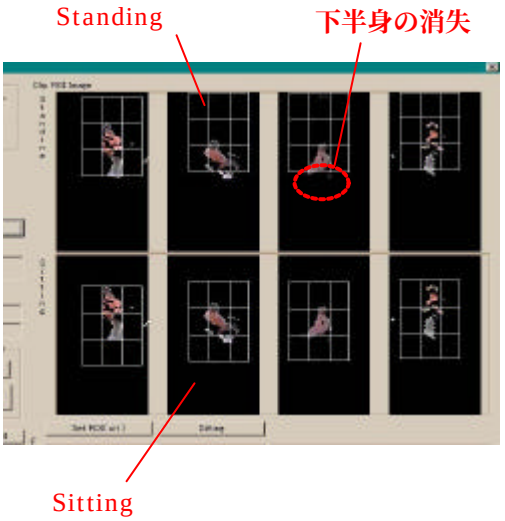

図 15：２種類の領域設定による認識

５．おわりに

人間のジェスチャーを本質的に理解するための新しいモデルとして，View and Motion-based Aspect Model (VAMBAM)を提案した．また，このモデルを用いた具体的なシステムとしてジェスチャー認識システムを構築し，提案したモデルにより人物の位置および回転に依存せずに DOVS 環境下でジェスチャーの認識が実現可能であることを示した．また，実時間人間追跡システムと時空間連続 DP（STCDP）を用いることによって，本手法の有効性を示した．さらに，本システムをより一般化させるために，椅子や机等を配置した DOVS 環境での人間の行動認識を実現した．

今後は，さらにモデルを追加することで認識可能な動作を増加を行い，次には人間が机上の物体を持ち上げる行動の認識や，握手や抱擁等の人間同士の相互作用の認識を行い，最後には人間が複数集まって形成する群衆レベルの認識を実現することを目標とする．

参考文献

[1] 八木康史，“実時間全方位視覚センサ”，日本ロボット学会誌，Vol.13，No3，1995.

[2] 天本直弘，藤井明宏，“画像処理技術による障害物検出と移動物体追跡方法”，電子情報通信学会誌，Vol.J81－A，No.4，1998.

[3] 長井敦，久野義徳，白井良明：，複雑変動背景下における移動物体の検出”，電子情報通信学会論文誌，Vol.J80－D2，No.5，1997.

[4] Y. Kitamura and M. Yachida, Three- dimensional data acquisition by trinocular vision, Advanced Robotics, Vol.4,No.1,pp.29-42,1990.

[5] 十河 卓司，石黒 浩，モーハン M. トリベディ，“複数の全方位視覚センサによる実時間人間追跡システム”，電子情報通信学会論文誌，2000.

[6] 岡 隆一，“連続 DP を用いた連続音声認識”，音響学会音声研資料，S78-20，pp.145-152 (1978-06).

[7] 西村 拓一，向井 理朗，野崎 俊介，岡 隆一，“低解像度特徴を用いた複数人物によるジェスチャの単一画像からのスポッティング認識”，信学論(D-Ⅱ),Vol. J80-D-Ⅱ，No.6，pp.1563-1570,1997.

[8] H. Ishiguro, “Development of Low-Cost Compact Omnidirectional Vision Sensors and Their Applications,” International Conference on Information Systems, Analysis and Synthesis, pp. 433-439, 1998.

[9] D. Marr, Vision: A computational investigation into the human representation and processing of visual information, Freeman, W.H.and Company, 1982.

[10] V. I. Pavlovic, R. Sharma and T. S. Huang, Visual Interpretation of hand gestures for human -computer interaction: A review, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 677-695, 1997.