ジェスチャと音声による屋内型飛行ロボットの対話的航行

中村 竜二† 細井 一弘† 杉本 雅則†
†東京大学大学院 新領域創成科学研究科 基盤情報学専攻
〒277-0882 千葉県柏市柏の葉 5-1-5 基盤棟 7G6
E-mail: †{nakamura, hosoi, sugi}@itl.k.u-tokyo.ac.jp

あらまし 本研究では，人間とのインタラクションに基づく屋内型の小型飛行ロボットの自律航行システムを提案する．ペイロードが小さい飛行ロボットには，マイク内蔵の単眼カメラのみを搭載した．3 次元復元 (因子分解法) を用いたポインティングジェスチャー認識および音声認識により，飛行ロボットはユーザの指定した方向，場所に移動する．外界の情報をカメラで得られない場合は，音声情報で補完するなど，画像，音声情報を統合することで，飛行ロボットが効率よく航行できることを示せた．
キーワード blimp, gesture recognition, single camera

Interactive Navigation of a Flying Robot Using Gesture Recognition and Voice Recognition

Ryuji NAKAMURA† Kazuhiro HOSOI† Masanori SUGIMOTO†
†Dept. Frontier Informatics, Graduate School of Frontier Sciences, The University of Tokyo
7G6, 7th Floor, Frontier Sciences Bldg., 5-1-5 Kashiwanoha, Kashiwa-shi, Chiba
E-mail: †{nakamura, hosoi, sugi}@itl.k.u-tokyo.ac.jp

Abstract In this paper, we propose an aunotomous navigation system for a flying robot based on human robot interaction. The payload of a blimp that we use is small, thus we mounted a single wireless camera with a microphone only. Once a user points at a target or a direction, the flying robot will recognize the gesture and move toward the target or direction in a 3D space. When the blimp cannot retrieve camera images, it utilizes user's voice information captured from the camera's on-board microphone. By integrating the visual and auditory information, a robust and effective navigation on the blimp interaction with a user was proved to be possible through experiments.
Keyword blimp, gesture recognition, single camera

1 はじめに

近年，ロボティック技術の発達に伴い，監視，ガイドなど，多岐に渡る用途のためのロボットが，われわれの身の回りで実現されている．

本稿では，それらのロボットのうち，移動の自由度が高く，広い視野を持つ飛行ロボットに注目し，人間の直感的な操作により，飛行ロボットを目的地に導くシステムの構築を目指す．

人間とロボットとが対話的にやり取りを行い，ロボットを目的地に導くような手法は以前から広く研究がなされてきた [1]．しかし，そのほとんどは，制約の少ない地上ロボットのナビゲーションを目的としており，飛行ロボットを対象にした手法は，ほとんど研究がなされていない．

そこで，本研究では，マイクの内蔵された単眼カメラのみをセンシングデバイスとして搭載することで，人間のジェスチャー認識と音声認識を利用した，自律飛行ロボットのマルチモーダルなナビゲーションを行うことを考える．これによって，大型展示場でのナビゲーションや，監視など，人間が直接その場の情報を得られないような場所におけるアプリケーションが実現できる．

ここで，ロボットのナビゲーションにおいて，ロボットに単眼のカメラを固定して設置した場合，目標物を見つける，見失ってから復帰する，といったことが困難である．そのため，本稿で扱うナビゲーションでは，

- 人間が行ったジェスチャーに従い，飛行ロボットが目的地を設定，航行する
- 航行中，人間が意図した目的地と異なる方向へロボットが向かった場合には，人間が音声によってロボットに指示を出すことで，航路を訂正する

というシナリオを想定している．

以下，第 2 節で今回提案する手法の中で用いている技術の概要について述べ，3 節でそれらの実装と実験結果について示した後，4 節に結論をまとめる．

2 提案手法

本稿では，ジェスチャーの中でも特にポインティングジェスチャーに注目し，人間が指し示す方向に飛行ロボットを航行させることを目的とした．図 1 に本システムの概要を示す．

本システムでは，図に示した手法を用いることで，人間は特別な機器を身に付けることなく，ロボットのナビゲーションをすることができる．また，同時に，ポインティングジェスチャーと音声によるナビゲーションという，人間の直感的な動作による操作を実現している．以下に，それぞれの段階において用いられている手法の概要を述べる．

ポインティングジェスチャーの認識に際しては，様々な手法が既に提案されている．しかし，飛行船に搭載したカメラでは，画像中の人間の位置や大きさが一定ではないという特有の性質がある．このため，今回の実装に当たっては，人間の顔および手の 3 次元位置を計算することで，人間の指し示す方向の推定を行った．

本研究では，まず，マハラノビス距離を用いた肌色認識によって，顔及び手の領域を抽出する．ここで，マハラノビス距離とは，多変量解析の一手法として用いられているものであり，複数系列のデータの相関から計算された類似度である．

こうして抽出された領域に対して，因子分解法 [7] と呼ばれる手法を適用することで，その 3 次元情報を再構築する．因子分解法とは，与えられた画像のシーケンスから，映っている物体の形状やカメラの動きを復元する手法であり，SfM(Structure from Motion)[2] と呼ばれる技法の一つである．

因子分解法により求められた 3 次元情報から，人間の指し示す方向を同定した後は，指し示されたマーカを推定し，その方向へ向かって飛行ロボットを航行させる．この際には，オプティカルフローを用いることで回転角度及び縦方向の移動を検知し，指し示した方向まで回転すると同時に，上下方向の移動を抑えている．

目標とする方向まで回転した際および，目標とするマーカが一旦カメラに写った際には，飛行船は，そのマーカに向かって航行を行う．この際には，飛行船の制御においてPD 制御を行うことで，安定した航行を実現している．

また，全体の航行を通して，音声認識による人間とのインタラクションを加えることで，より効率的な航行を実現した．音声認識に際しては，フリーソフトとして提供されている大語彙連続音声認識システム Julius[10] を用いている．

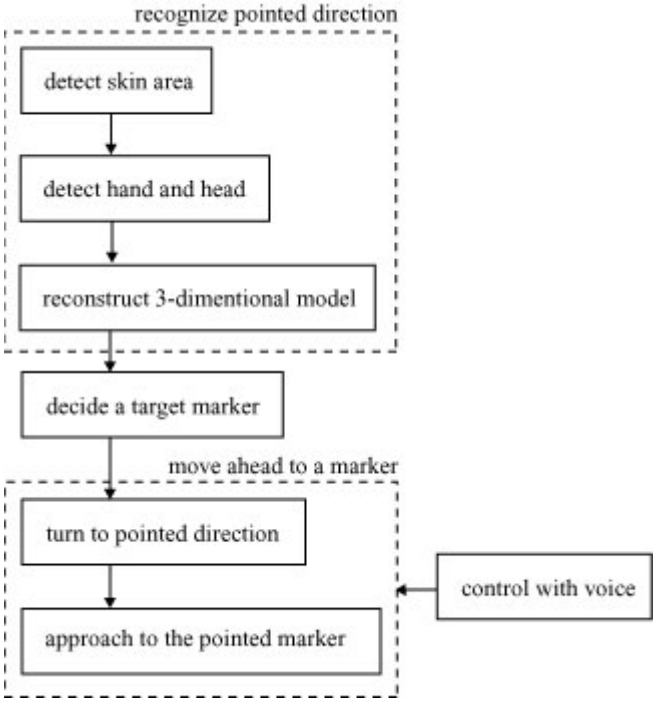

図 1: 本システムの概要

以下に，それぞれの段階において用いている手法の詳細を示す．

2.1 マハラノビス距離を利用した肌色認識

本研究では，現在数多くの手法が提案されている肌色認識の手法中でも，特に Wakamura ら [8] の提案する，マハラノビス距離を用いた肌色認識の手法を利用する．この手法では，光源の影響を受けにくい色相と彩度を用いてベクトルを構成し，そのベクトルからマハラノビス距離を求めている．

一般に，マハラノビス距離を利用した肌色領域検出は，大きく 4 つの段階に分かれて，処理が行われる．

1. カメラから得られた画像を HSV 色空間に変換し，その中でも色相と彩度のみを特徴空間として肌色のクラスタを生成する
2. 肌色のクラスタの平均ベクトル $M = [M_H M_S]^T$ および共分散行列 V を計算する．ここで，共分散行列 V は以下で表される．
$$V = \begin{bmatrix} \sigma_X & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y \end{bmatrix}$$
3. あらかじめ得られている平均ベクトル M 及び共分散行列 V に対し，計測された特徴量ベクトル X とのマハラノビス距離を以下のように与える．
$$d_M^2 = (X - M)^T V^{-1} (X - M)$$

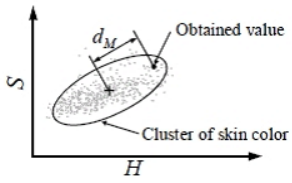

図 2: マハラノビス距離とその分布

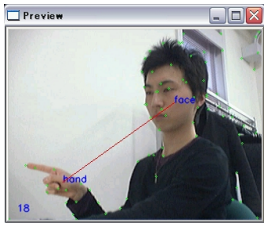

図 3: 顔と手の検出結果

4. このとき，マハラノビス距離 d_M の値が設定した閾値の範囲内にある画素を肌色領域とする (図 2)

本研究では，あらかじめ用意したテストデータに対して肌色領域となる部分を抽出し，そのデータを元に平均ベクトル M，共分散行列 V および閾値を定めた．

2.2 顔と手の位置の推定

顔と手の領域の位置の推定においては，まず一画面中に一人の人間のみが写っているという仮定を置いている．まず，前節のようにして求められた肌色領域に関して、収縮処理及び膨張処理を行い，孤立点除去および近傍領域の連結を行う．次に，肌色領域と判別された部分をラベリングし，ラベリングされた領域に関して領域のサイズが最も大きい部分を顔領域，次に大きい部分を手領域としている (図 3)．

このようにして得られた顔領域および手領域の中心を用いて，ポインティングジェスチャーで指し示されている方向を推定する．ここでは，手領域の重心と顔領域の重心とを結ぶ線が，指し示されている方向に近似されることを仮定している [6]．

2.3 因子分解法を用いた 3 次元復元

因子分解法では，与えられた画像のシーケンスから，画像中に共通して映っている物体の 3 次元形状を再構築する．ここでは，まず最初のフレームに対し，特徴点を抽出し，それをすべてのフレームに対してトラッキングし [5]，すべてのフレームに共通して含まれる特徴点を行列に格納する．ここで，シーケンスのフレーム数を F，対応する特徴点の個数を P とおく．すると，すべての特徴点をまとめた行列として以下が得られる．

$$W = \begin{bmatrix} u_{11} & \cdots & u_{1P} \\ \vdots & \ddots & \vdots \\ u_{F1} & \cdots & u_{FP} \\ v_{11} & \cdots & v_{1P} \\ \vdots & \ddots & \vdots \\ v_{F1} & \cdots & v_{FP} \end{bmatrix}$$

ここで，f フレーム目，p 番目の特徴点のフレーム中の水平座標を u_{fp}，垂直座標を v_{fp} とする．1 つのフレーム内のすべての特徴点に関して，その平均を水平座標，垂直座標それぞれに対して求め，そのフレーム中の点の座標系を求めた平均値を基準とした座標系に変換する．

$$a_f = \frac{1}{P}\sum_{p=1}^{P} u_{fp},\ b_f = \frac{1}{P}\sum_{p=1}^{P} v_{fp}$$

このとき，フレーム f における点 P の座標系を平均値 a_f, b_f を基準とした座標系に変換する．これによって u_{fp}, v_{fp} は以下のように変換される．

$$\tilde{u}_{fp} = u_{fp} - a_f$$
$$\tilde{v}_{fp} = v_{fp} - b_f$$

変換後の W を $\tilde{W}$ とおくと，これは，カメラの回転ベクトル R と物体の形状ベクトル S を用いて以下のように表される．

$$\tilde{W} = RS$$

ここで，それぞれのフレームにおいてワールド座標系における基底ベクトルを $\mathbf{i}_f, \mathbf{j}_f$($\mathbf{i}_f$ は水平方向，$\mathbf{j}_f$ は垂直方向) とおくと，R は
$R = [\mathbf{i}_1\ \ldots\ \mathbf{i}_F\ \mathbf{j}_1\ \ldots\ \mathbf{j}_F]^T$ で表される．こうして定めた $\tilde{W}$ に関して，今回のアルゴリズムでは以下の操作を行う．

1. $\tilde{W}$ を以下で与えられる式に従って特異値分解する
$$\tilde{W} = O_1 \Sigma O_2$$
ここで，O_1, O_2 は次の式を満たす行列であり，Σ は $P \times P$ の対角行列である (I は単位行列)
$$O_1^T O_1 = O_2^T O_2 = O_2 O_2^T = I$$

2. 以下のように $\hat{R}, \hat{S}$ を定義する
$$\hat{R} = O_1'(\Sigma')^{1/2},\ \hat{S} = (\Sigma')^{1/2} O_2'$$
ただし，O_1', O_2', Σ' は以下のように分解された行列であり，それぞれ $2F \times 3, 3 \times P, 3 \times 3$ の大きさを持つ
$$O_1 = \begin{bmatrix} O_1' & O_1'' \end{bmatrix},\ \Sigma = \begin{bmatrix} \Sigma' & 0 \\ 0 & \Sigma'' \end{bmatrix},\ O_2 = \begin{bmatrix} O_2' \\ O_2'' \end{bmatrix}$$
また，これらを用いて $\tilde{W}$ は
$$\tilde{W} = O_1 \Sigma O_2 = O_1' \Sigma' O_2' + O_1'' \Sigma'' O_2''$$
と表される

表 1: Julius によって認識する語彙

words (direction)	右，左，上，下，進め，戻れ，止まれ
words (manipulation)	始め，終わり
words (adjective)	もっと
phrases	adjective + direction

3. 基底ベクトル $\mathbf{i}_f, \mathbf{j}_f$ を用いて，次の条件を満たす行列 Q を定める

$$\hat{\mathbf{i}}_f^T Q Q^T \hat{\mathbf{i}}_f^T = 1,\ \hat{\mathbf{j}}_f^T Q Q^T \hat{\mathbf{j}}_f^T = 1,\ \hat{\mathbf{i}}_f^T Q Q^T \hat{\mathbf{j}}_f^T = 0$$

4. 求められた Q を用いて以下の式から R, S を求める

$$R = \hat{R}Q,\ S = Q^{-1}\hat{S}$$

以上の演算によってカメラの回転とカメラに写っている物体の 3 次元形状を再構築することができる．

本実験では，顔領域の重心として得られた点および手領域の重心として得られた点の 3 次元形状の構築を行うことで，計算量の軽減を図った．

2.4 Julius による音声認識

本稿で用いられている Julius[10] は，統計言語モデルである単語 N-gram を用いて音声認識を行う．テキストコーパスから学習された単語 3-gram を用いて，大語彙の汎用音声認識 (書き下し/ディクテーション) を行うことができる．また，Julius はユーザが定義した記述文法に基づいた認識を行うこともでき，本実験では，専用の文法を記述し，それに基づいた音声認識を行うことで，精度の高い認識を実現している．

Julius を用いた音声認識においては，本プログラム中で用いている 10 個の語彙 (表 1) を登録し，それらの入力があった際には，画像から得られた情報に基づく制御よりも優先度の高い操作として，音声認識結果から生成した制御信号を送信している．

3 実装と実験結果

本節では，まず，それぞれの段階 (肌色認識，3 次元復元，回転角度計算，音声認識) における結果を示した後，プログラム全体における精度を，画像のみによる航行と，画像と音声を用いた航行の 2 つに関して比較，評価する．

ここで，飛行船の航行に際しては，Mobile Airships & Blimps 社 [9] の飛行船を用いており，航行に際しては，左右方向に回転するためのプロペラ及び上下方向に航行するためのプロペラをオン，オフすることで飛行船を動かしている．このとき，モータを駆動させるための最小の荷電圧の時間が 80[ms] であるため，このときの最小の回転角度が 14.3[deg] 程度であることが分かっている．

図 4: 肌色認識の結果

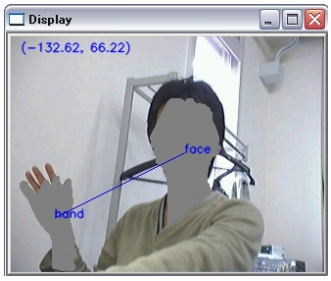

図 5: 顔・手認識の結果

3.1 肌色領域の認識結果

本実験では，あらかじめ用意した 20 枚のテストデータに対して肌色領域となる部分を抽出し，そのデータを元に平均ベクトル M および共分散行列 V を以下のように求めた．

$$M = \begin{bmatrix} 6.454736842 & 74.80473684 \end{bmatrix}$$

$$V = \begin{bmatrix} 3864.4 & 233.2 \\ 233.2 & 1933.0 \end{bmatrix}$$

また，これに対して閾値を以下のように定めた．閾値の設定においても先の 20 枚のテストデータを用い，特に手領域が完全にカバーできる範囲での値を閾値と定めた．

$$Threshold \begin{pmatrix} max & min \end{pmatrix} = \begin{bmatrix} 539.7831127 & 0.818265778 \end{bmatrix}$$

ここで，これらの平均ベクトル M，共分散行列 V および閾値による肌色領域検出の精度を，テストデータを除く，100 サンプルに対してテストすると，87 %の精度で肌色領域が検出されたことが確認された．ここで，検出が適切に行われたかどうかの判断は，人間の肌色の領域が完全に検出されており，かつ肌色領域と判断された領域のうち 80 %以上が肌色領域である場合を正例とすることで行った (図 4)．この図では，図中の灰色で塗りつぶされている領域が肌色として認識された領域である．

本手法で負例となる場合は，撮影環境があまりに暗く，また，光源も通常の部屋のような状況とは大きく異なる場合が多いことが確認された．また，背景の色成分が人間の肌に似ているような場合 (木目調の机など) において，背景も肌色領域として認識されてしまっていることも確認された．

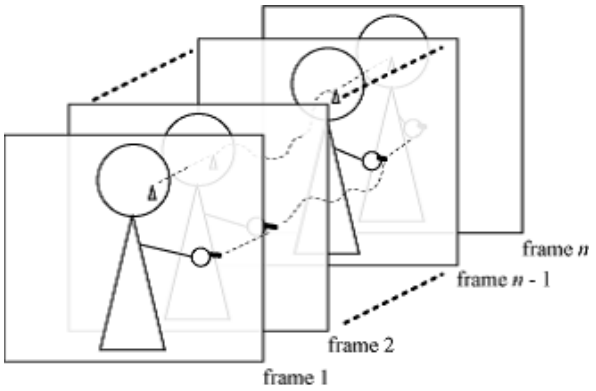

図 6: 因子分解法の精度評価のために用いたデータの模式図

表 2: 3 次元復元の結果

単位 [deg]

dimension of motion	ideal angle	calculated angle	difference
1D	28.898	26.777	2.121
2D	59.864	62.176	2.312
3D	50.081	56.31	6.229

3.2 顔と手の位置の推定の結果

領域の広さによる顔と手の認識に関して，先に実験を行った 100 サンプルの中から人間が一人のみ含まれており，かつその人間の手と顔が含まれている 20 サンプルに対し，実験を行った．その結果，60 %の精度で検出がなされることが確認された．ここで，検出が適切に行われたかどうかの判断は，顔及び手の領域の重心を求められた場合を正例としている (図 5).

ここで，負例となった場合について注目すると，今回，ポインティングジェスチャーの検出に焦点を絞ったアルゴリズムを設計する上で，顔と手の重心の距離が顔の領域の広さに対してある閾値よりも大きい場合のみ，顔領域と手領域を検出するという手法を取っているために，サンプルに写っている被写体の状況 (たとえば顔の近傍に手があるような場合) によって精度が下がってしまったことが考えられる．

3.3 3 次元形状の復元結果

本実験では，あらかじめ位置関係の分かっている 2 点に関して，視点を 1 次元，2 次元，3 次元で動かしたフレームシーケンスにおける，対応する点のスクリーン座標系での位置をデータとして与え (図 6 に，与えたデータの模式図を示す)，それから計算された 3 次元形状復元結果の精度を評価した．

表 2 にその結果を示す．ここでは，顔と手の中心点の代わりに定めた 2 点が視点に対してなす Yaw 角に関して，モデル上での角度 [deg] と，因子分解法を用いて計算された角度 [deg] との比較を示している．この結果で生じている角度の誤差は，計算上での誤差及び，モデルから特徴点を抽出する上での誤差によって生じていると考えられる．これは，今回，実際のカメラで用いる，360 × 240 の解像度の画面上で動きをとっているため，1 ピクセルの違いによって生じる誤差が大きくなってしまうためである．ここで，3 次元での動きの際に，1 次元，2 次元に比べて約 3 倍程度の誤差が生じているが，これは，与えたモデルの動きを，実際にカメラがとりうる動きよりもかなり大きくとったためである．これから，視点の移動の次元に依らず，計算された角度誤差は，飛行船のとりうる最小の回転角度に比べて，十分小さいと言える．

表 3: 回転角度の計算結果

$n = 20$, static condition: $\sigma = 2.22$, moving condition: $\sigma = 1.37$, 単位 [deg]

	static condition	moving condition
actual angle	0	10.24
calculated angle	0.46	11.31
difference	0.46	1.07

3.4 回転角度の計算

本手法では，飛行船の回転角度を計算するために，オプティカルフロー [3] を用いている．オプティカルフローの使用に際しては，Tomasi ら [5] の手法によって抽出した特徴点を，以降 30 フレームに関して追跡し，それらの点の動きから，フローを計算している．今回用いたオプティカルフローの計算の精度を表 3 に示す．ここでは，静止状態及び動いている状態における，実際の回転角度 [deg] と計算された角度 [deg]，およびびそれらの差を記載している．この結果から，オプティカルフローを用いた回転角度の計算による誤差は，先に記載した飛行船のとりうる回転角度に比べて，無視できる程度に小さいと考えられる．

3.5 Julius による音声認識

Julius による音声認識結果を表 4 に示す．ここで，ここに示した結果は，無線カメラに内蔵されたマイクを用いてそれぞれ 100 回ずつ試行した場合の認識結果であり，ノイズの全くない状態 (USB や有線のマイク) の場合には，ほぼ 100 %の認識率を実現している．この結果において，”左”や”下”，”進め”など，語頭がサ行で始まる場合には，カメラのノイズのために認識率が大変低くなっていることが分かった．また，その単語のみでは認識精度が悪い場合 (”左”や”下”) においても，語頭に”もっと”をつけることで，平均して約 92 %の精度が実現できることが分かった．

3.6 プログラム全体としての評価

ここでは，システム全体を通して，画像のみを用いて航行した場合と，画像による航行に，音声による割り込

表 4: 音声認識の結果

$n = 100$, 単位 [%]

order	recognition rate	recognition rate with "もっと"
右	90	20
左	10	60
上	80	80
下	10	100
進め	0	30
戻れ	100	100
止まれ	100	-
始め	70	-
終わり	100	-

表 5: 画像のみを用いた航行と画像と音声を用いた航行との航行時間の比較

$n = 10$, $\sigma = 29.2$, 単位 [sec]

	total time	turning time	approaching time
vision only	127.9	86.9	41.0
vision + voice	102.4	77.6	24.8

みを加えた場合との比較を行う．表 5 にその結果を示す．ここで，turning time は，回転して目標物の方を向くまでの時間 [sec] を示し，approaching time は目標物を見つけてから，ある距離まで近づくまでの時間 [sec] を示している．この結果から，飛行船が目的の場所にたどり着くまでの時間は，画像のみを用いた場合に比べて，画像と音声を用いた場合の方が，短縮されることが分かる．これは，無線カメラのみを用いた場合には，無線カメラからの信号にノイズ乗った際，制御ができなくなる一方で，音声による操作を加えた場合は，ノイズが乗った際にも飛行船に制御信号を送り続けることができるため，より円滑な航行が実現できるためであると考えられる．

4 まとめと今後の課題

本稿ではジェスチャーの認識による飛行ロボットの自律航行に際して，肌色認識，顔・手の認識および因子分解法を用いた肌色領域の 3 次元形状の復元を行い，それを利用したジェスチャー認識の手法を提案，実装した．また，それを利用して，飛行船航行を行い，画像処理のみによる航行と，画像処理に加えて音声認識を用いた場合の航行との比較を行った．

今後の課題としては，以下が考えられる．

- 無線カメラのノイズに起因する航行時間の遅延が生じているため，ノイズ耐性のあるアルゴリズムを提案する
- 肌色認識等において，制約を多く設けているため，よりロバストな手法を考案する

また，今後は音声だけでなく，他の様々な手法を併用することで，よりインタラクティブで直感的な飛行船操作の手法を検討していく予定である．

参考文献

[1] Desouza, G.N., Kak, A.C., "Vision for mobile robot navigation: a survey," IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 24, No. 2, February 2002, pp.237-267.

[2] Dellaert, F., Seitz, C., Thorpe, C., Thrun, S., "Structure from Motion without Correspondence," IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'00), June, 2000, pp.557-564.

[3] Gutierrez, M.A., Moura, L., Melo, C.P., Alens, N., "Computing optical flow in cardiac images for 3D motion analysis," Proceedings of Computers in Cardiology 1993, September 1993, pp.37-40.

[4] Kehl, R., Gool, L.V., "Real-Time Pointing Gesture Recognition for an Immersive Environment," Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition (FGR '04), May 2004, pp.577-582.

[5] Shi, J, Tomasi, C, "Good features to track," Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1994, pp.593-600.

[6] Stiefelhagen, R., Fugen, C., Gieselmann, R., Holzapfel, H., Nickel, K., Waibel, A., "Natural human-robot interaction using speech, head pose and gestures," Proceedings of IEEE Intelligent Robots and Systems, 2004. (IROS 2004), Volume 3, 28 Sept.-2 Oct, 2004, pp.2422-2427.

[7] Tomasi, C., Kanade, T., "Shape and Motion from Image Streams: a Factorization Method, Full Report on the Orthographic Case," http://vision.stanford.edu/public/publication/tomasi/tomasiTr92Text.pdf, 1992. Letzter Zugriff: 12.03.2005.

[8] Wakamura, N., Suzuki, K., Umeda, K., "Construction of an Intelligent Room Using Intuitive Gesture Recognition," 日本機械学会ロボティクス・メカトロニクス講演会'05 講演論文集，2A1-N-051, 2005.

[9] http://www.blimpguys.com/

[10] http://julius.sourceforge.jp/