複合コミュニティ空間のための人物動作理解と人物映像加工

Human behavior understanding and human image processing for MIXED COMMUNITY SPACE

中村 裕一 † 北原 格 † 坂上 勝彦 ‡ 大田 友一 †

Yuichi NAKAMURA Itaru KITAHARA Katsuhiko SAKAUE Yuichi OHTA

† 筑波大学 機能工学系

† IEMS University of Tsukuba

‡ 電子技術総合研究所

‡ Electrotechnical Laboratory

1 はじめに

「複合コミュニティ空間」とは，現実世界と仮想世界が融合した複合現実の感覚を複数の人間が共有するための，コンピュータによって形成される人工空間である．すなわち，複合コミュニティ空間では，複数の人間同士が現実世界で行っている視覚的な情報交換と現実世界には無い新たな視覚情報の両者を同時に共有する．我々の研究では，このような人工空間を自然なかたちで実現するための新しいパターン認識・理解の枠組みを追求することを目的としている [1]．

2 複合コミュニティ空間の課題

現在の複合現実感技術では，観測・通信・提示の種々の問題により，現実世界で一緒に行動している人間同士のコミュニケーションが阻害される．例えば，頭部装着型ディスプレイ (HMD) を着用すると，一緒に行動している人間同士の視線や表情が読み取りにくくなる．

そのため，人間を中心とした実時間対話環境を実現するためには，複合現実感を実現するために失われる情報，伝わりにくい情報を復元すること，必要に応じてそれらを強調・加工して伝えることが必要となる．そのために我々が取り組んでいる具体的な課題を以下にあげる．

- 互いの顔や動作を視認するための視覚支援
- 複数人による注意や注目の共有のための視覚支援
- 互いを含む周囲の状況を把握するための視覚支援
- 必要な情報を適所で提示するための視覚支援

本研究では，上記の個々の機能を実現するだけでなく，これらの機能を統合した複合コミュニティ空間のプロトタイプシステムを構築し，その評価を行うことを目指している．

3 顔を視認するための視覚支援

顔は人間にとって最も親しみのある対象の 1 つであり，特に視線，表情は音声では表現できない情報を伝える．我々の研究では，このような情報伝達形態を阻害しないように，HMD をかけた人物の顔に，HMD をかけていない状態の顔画像を重ね合わせて表示する．

3.1 任意視線の顔画像生成:

HMD 装着者の視線方向がわかっている場合[1]，その方向に対応する眼付近の顔画像を生成し，撮影された映像に重ねる．図 1 にシステムの全体像を示す．IBR (Image Based Rendering) の手法を用いており，視線の異なる複数枚の顔画像があれば，任意の視線をもった顔画像を生成することができる．図 2 に，生成された顔画像の一部を示す．左が生成した任意視線の顔画像であり，右が実際に同じ方向を見た時の顔画像である．

このような画像を生成する手法として，単に眼球を移動しただけでは不自然な顔画像になってしまう．そのため，眼を開いている画像と閉じている画像から中割り法によって中間画像を生成し，まぶたの動きを再現する．これらを統合することでより自然な任意視線の顔画像を生成する．視線移動は眼球の回転という 3 次元運動により起こる．しかし眼球は黒目と白目という比較的単純な構造を持っており，また眼球すべてが見えているわけではない．このことから，眼球の回転運動を 2 次元の平行移動で近似することは十分に可能であると思われる．

まばたきによる眼の周辺部分の動きは比較的単純であるが，そのテクスチャ変化は非常に複雑である．特に，まぶたは眼を閉じたときにはすべて露出されているが，目を開いた時には (一重や二重といった) しわに折り畳まれて見えなくなってしまう．しかし，IBR 法によれば，比較的容易に自然なまぶたの動きの再現が可能と考えられる．

本手法を用いることで，視線の異なる複数枚の顔画像が

[1]本研究で，視線検出機能付き HMD を試作した．

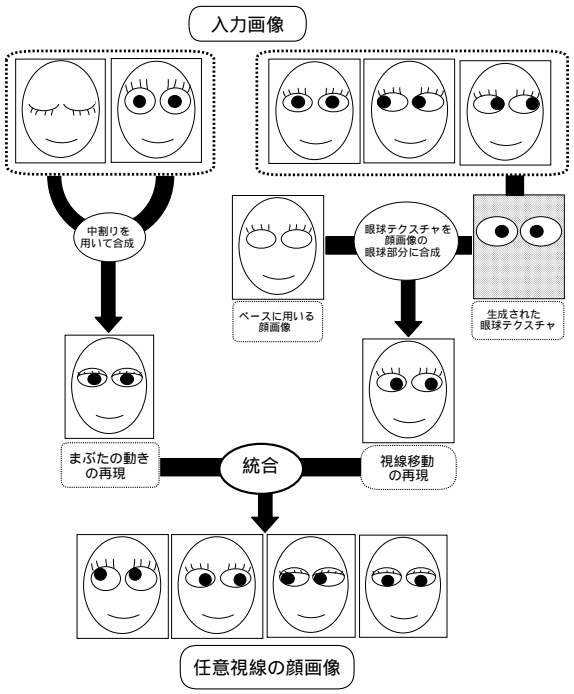
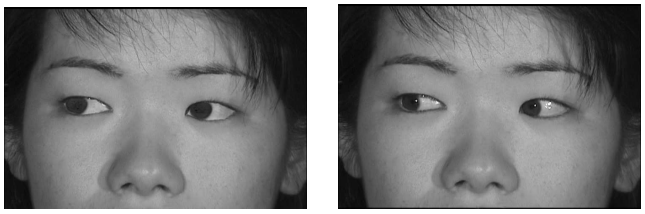


図 1: 任意視線の顔画像生成システム



(a) 合成画像　　(b) 実写画像

図 2: 任意視線の顔画像生成

あれば，任意の視線をもった顔画像を生成することができる．また，眼を開いた顔画像と眼を閉じている顔画像の 2 枚あれば，まばたきなどまぶたの動きを再現することができる．さらに，これらを組み合わせることで実際の人間のように視線移動に沿った自然な任意視線の顔画像を生成することができる．

眼球テクスチャの生成

入力画像の眼球部分を単純に平行移動しただけでは，生成された顔画像においてテクスチャの欠落が生じる．これを補うため人間の眼球全体を再現した眼球テクスチャを生成する．理想的には眼球すべてを覆うテクスチャを作れば良いが，実際の視線移動では眼球がすべてみえるわけではないので，視線移動によって見える部分だけを眼球テクスチャとして作成すれば十分である．眼球は黒目と白目の境界がはっきりしているので，眼球全体で考えると白目部分内での色の変化は少ない．このため白目のテクスチャ同士

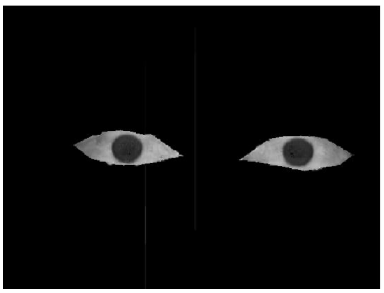


図 3: 眼球テクスチャの例

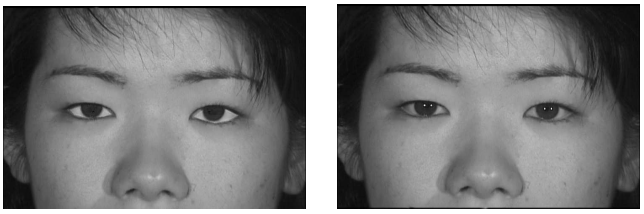
が重なり合ってもその境界付近に生じる極端な色の変化は起こりにくい．このことから，あらかじめ視線の異なる複数枚の入力画像から眼球部分のテクスチャを切り出し，黒目位置を基準にして貼り合わせることで眼球テクスチャを作成する．生成された眼球テクスチャの例を図 3 に示す．

視線移動の再現

眼球とまぶたを動かすことによって，視線移動を再現する．眼球テクスチャを平行移動し，任意視線を表現したい顔画像の眼球部分に合成することで任意視線を表現する．しかし，生成する顔画像と位置基準に用いた顔画像の大きさや向きが必ずしも同じでないので，生成する顔画像に合わせて眼球テクスチャに回転・拡大・縮小などの変形を施す必要がある．また，人間のまぶたは視線移動と非常に密接な関係があり，視線を移動するとそれに伴いまぶたなど眼のまわりも動く．具体的には視線を下に落とした時には，上を向いた時の半分ぐらいまでまぶたが下がっている．そのため，眼球テクスチャの平行移動による視線移動とともにまぶたの動きを再現することは重要である．本研究では，眼を開いた画像と閉じた画像を重み付け平均を用いてマッピングすることで，しわも表現した自然な画像を生成する．

実験方法

任意視線の顔映像を生成する実験を行った．合成対象の顔画像に合わせて回転・拡大・縮小した眼球テクスチャを，表現したい視線に合うように平行移動し，顔画像の眼球の部分にはめ込むことで任意視線の顔画像を生成した．顔表面上に配置された特徴点のうち，両目尻の 2 点を回転成分を求めるための基準点として用い，両目の中間点，右目尻，鼻の頭の 3 点を拡大・縮小率を求めるための基準点として用いた．図 4 に，生成された顔画像の一部を示す．左が生成した任意視線の顔画像であり，右が実際に同じ方向を見た時の顔画像である．実験により，少数の実画像から生成された一枚の眼球テクスチャを用いて，様々な視線の顔画像を生成できることが確認できた．



(a) 合成画像　　(b) 実写画像

図 4: 任意視線の顔画像生成

3.2　任意方向・表情・口形状の顔画像生成

任意の方向・表情・口の動きを持った顔映像を生成し，それを現実の映像中の人物像と合成する．これによって，映像中の人物の顔を，同じ姿勢・表情・発話の状態で，他の顔にすり替えることができる．

その概要を図 5 に示す [3]．2 枚の顔画像間で顔上に配置した特徴点が対応付けられているとき，任意方向の顔画像上の特徴点座標は，2 枚の画像上での特徴点座標の線形結合で表現できる．マッピングするテクスチャは生成したい顔の向きと近い方向の入力画像から優先的に獲得し，重みづけ平均をとることによって任意方向での自然な顔画像の生成を可能とする．同一方向から撮影された，異なる表情・口形状を持ったいくつかの画像があれば，これらの画像上で観測される特徴点の線形結合によって，任意の表情・口の動きでの特徴点座標を算出することができる．これらの特徴点座標をもとに生成したい表情と口の動きに近いテクスチャを優先的にマッピングすることで，自然な表情・口の動きをもった顔画像が生成できる．

合成の対象となる元映像をここではサンプル映像，上書きしたい人物を撮影した静止画像群を入力画像と呼ぶことにする．本実験では，サンプル映像として約１００フレーム（３秒間程度）の１シーンを用いた．また，笑った顔・驚いた顔・普通の顔など 6 表情と『あ・い・う・え・お・ん』の 6 つの口形で，合計１２種類の顔を，正面・左右から撮影した総計３６枚を入力画像として使用した．特徴点として，サンプル映像上では，鼻・額など 5 点，入力画像上ではマーキングによる８６点を抽出した．その例を図 6 に示す．

これらの特徴点をもとに表情・口形の異なる１２種類の入力画像のそれぞれについて，サンプル映像上 (図 7) の人物と同じ方向の顔画像を生成した．さらに，それらを組み合わせてサンプル映像と同じ表情・口の動きをもった顔 (図 8) を生成した．上書きする画像の表情と口の動きは，サンプル映像を参考に人間が手入力で指定した．

現在は，照明の効果も加えて，より自然な合成を可能とすること，また表情や口形状の指定を自動化すること，合成のリアルタイム化について研究を行っている段階である．

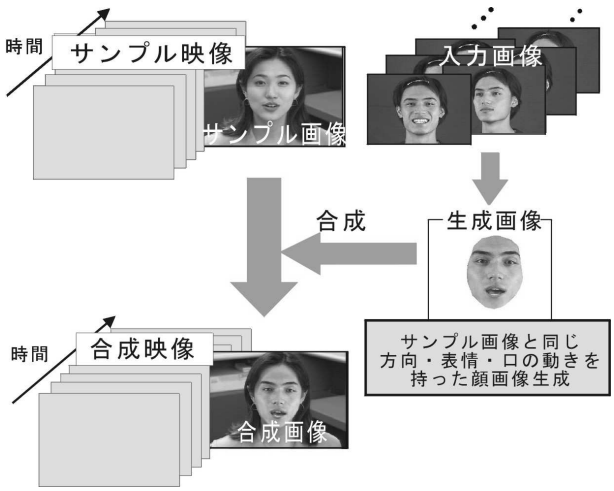


図 5: 任意方向・表情・口形状の顔映像生成

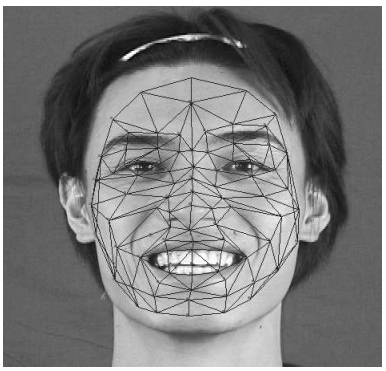


図 6: 特徴点と三角形パッチ

4　注意・注目の共有のための視覚支援

複合コミュニティ空間で，注意・注目を共有するために，各々の人間の動作・発話の意図を推定し，それによって各々に対する提示情報を制御する．

現在は，コミュニケーションにおいて特に重要である，指示動作・例示動作を対象とし，以下の項目について研究を進めている．

- 動作と発話の統合処理による，指示・例示動作の認識
- 指示対象や話題の対象になる人物や物体を撮影するためのカメラ制御
- 動作情報を用いた，提示情報の切り替え

そのために，動作の 3 次元データ (磁気センサから得られる)，発話データ (音声認識システムから得られる) を基に，オンラインで指示動作認識を行うシステムを構築した．これは，我々の研究で構築した人物動作データの撮影・蓄積システム [7] を基にし，動作認識をオンライン化したものである．認識に 1 ~ 2 秒程度の時間遅れがあるものの，簡単な指さし動作などは良好に抽出し，指先を重点的に提示することができる．その例を図 9 に示す．ここで

図 7: サンプル画像

図 8: 実験結果　上段：サンプル画像 下段：合成結果

は指示動作が行われた時に，その指示対象を最も大きく写した映像に切り替えて提示している．このような視覚支援によって，複数人での注目を共有する．

4.1　マルチモーダル・多視点データを用いた注目要素推定

注目要素の選択・提示

プレゼンテーションや協調的な作業では，動作の目的がはっきりとしており，話し手が伝えたい情報をよく推定することができる．ここで，話し手 (又は行為者) が相手の注意を向けさせようとしている対象を注目要素と考える．種々の注目要素のうち，特に重要な注目要素は，話者が陽に注目を促している状況，空間的な情報に注意が向けられている状況，特定の重要動作が行われている状況で生じると考える．

上記の考えに基づき，本研究で実際に対象とするのは以下の状況である．

指示:　物体を指差して，“これを...” と言う場合など，明らかに物体や場所を指し示している状況．

例示:　“ こんなふうに.. .” というように，身振りを使って情報を伝えようとする場合のように，例示動作 [10] の分類による象形動作，空間動作，活動動作が現われた状況．

(a) 指示動作

(b) 指示対象の提示

図 9: 指示動作による提示映像の切替

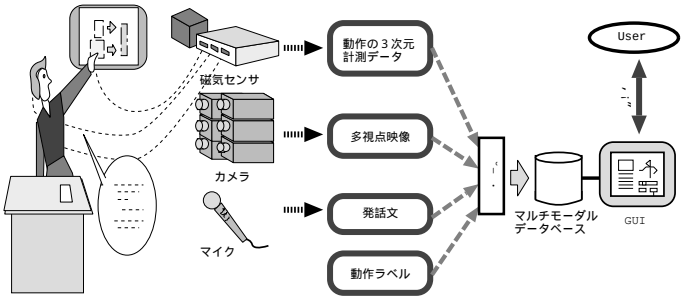


図 10: 多視点マルチモーダル映像取得システム

本研究では，これらの部分を認識し，強調して聞き手に提示する．

マルチモーダル・多視点データの必要性

話し手の言動から注目要素を推定するためには，話者の動き，発話を統合的に処理する必要がある．人間の動作には明確な規則がなく，状況，個人差に大きく影響されるからである．そのため，人間の動作理解を行うためには発話情報などの文脈情報の処理が必要であることを我々は指摘してきた [6]．また，発話情報には注目要素の存在を示唆する手がかりが多く含まれているが，実際に話し手が注目することを要求している空間的部分，時間的部分の特定は難しい．

上記の理由から，映像データの取得段階で動作，発話等の複数の情報を記録しておくことが必要となる．そこで，我々は図 10 のようなマルチモーダル・多視点映像取得システムを構築した．このシステムでは，映像 (画像，音声) とともに，発話，人物動作を関係付けたデータを記録する．

これは，以下のような構成からなっている．

多視点映像:　多数のカメラで，シーン全体，上半身，両手，指示物体，テーブル上の物体等を撮影する．カメラはホストコンピュータからズーム，パン，ティルトを制御される．これらのカメラによって，撮影対象を

図 11: 多視点映像の例

常に追跡しながら撮影を行うことができる．これにより，カメラは受け手の注意をひく可能性のある対象を高い解像度で撮影できる．この多視点画像の例を図 11 に示す．なお，撮影対象の追跡には，磁気センサで計測された身体各部の位置を利用している．

位置・動きデータ: 話し手の上半身の動作を磁気センサを利用して計測する．磁気センサは受信器の位置と方位角の 6 自由度を測定できる．一般的に，“動作” という単語が，手足の動きのようなデータを表す場合から，“切る” のような目的/効果を表す場合まで，幅広く用いられるため，他と区別する目的でこのデータを位置・動きデータと呼ぶことにする．

発話: 音声認識エンジン[2]を用いる．現段階では，連続音声を正確に認識することが困難なため，記録を残すためには，話し手の発話を人手で書き起こしたものを付加する．

このデータは，複数の視点から撮り流した映像を含むため膨大な量になるが，注目要素を推定し，それを基に伝送することにより，受手側と注意を共有をすることが可能になる．

マルチモーダル・多視点映像データベースの利用

マルチモーダル・多視点映像データから注目要素を推定する際に問題となるのが，前節で述べた人物の言動の多様性である．我々はこの問題を扱うために，料理やプレゼンテーションのマルチモーダル・多視点映像データを収集し，データベースを構築している [7]．本研究では，このデータに基づいて，注目要素推定処理のパラメータ設定やアルゴリズムの有効性評価等を行う．

データベースに蓄積されたデータを検索することにより，実際のプレゼンテーションにおける，動作 (動き)，動作 (目的)，発話，及びその相互関係を調べることができる．例えば，注目要素選択のためには，特定の目的を持った動作/発話の頻度，種類，共起性を用いることができる．

[2]現在は，ViaVoice を用いている．

表 1: 指示詞と現場指示動作との共起割合

指示詞	共起割合	指示詞	共起割合
あそこ	0.09	これ	0.22
あちら	0.05	こんな	0.20
あの	0.18	こういうふうに	0.25
こう	0.25	こんなふうに	0.26
ここ	0.13	そう	0.33
こちら	0.11	そこ	0.11
この	0.17	それ	0.09
このような	0.28	どう	0.25
このように	0.33	どこ	0.16

4.2 注目要素抽出とそれによる要約

注目対象を映像から抽出するためには，動作 (位置・動きデータ) の処理と発話の処理を相補的に用いることが必要となる．

言語的な手がかり

上記の注目要素を検出するための言語的な手がかりには，種々の単語や構文が考えられる．その中でも最も簡単で有効な手がかりは，指示詞や典型的な動詞である．

そのため，本研究ではまず，実際のプレゼンテーション場面における各単語の現れ方を調べ，それを基に各単語に注目フレーム検出のためのスコアを付与した．ただし単語として，形態素解析によって得られる形態素を用いる．これは，現在広く用いられている形態素解析システムの精度が非常に良いためである.

指示詞については，現場指示 (物体指示，場所指示)，例示の動作と各指示詞が共起した割合を求めた．ただし，各指示詞を含む文が話されている区間と動作が行われている区間が時間的に重複することを共起の条件とした．その結果を表 1 に示す．ここで与えられている割合は，

$$Q_i = N_c/N_a$$

である．ただし，N_c は単語が現場指示動作または例示動作と共起した回数，N_a はその単語の総出現回数である．

位置・動き特徴による手がかり

位置・動きデータは，指示動作の抽出だけでなく，指示動作が行われている時間的区間・場所の推定に用いる．本研究では以下の特徴量を用いる．

静止: 動作の開始直前，終了直後，指示動作の指示している瞬間は手が急停止し，静止している場合が多い．た

だし，連続的に複数の動作を行っている場合には，必ずしも静止点があるとは限らない．

腕の伸び: 指示動作では腕を伸ばすことによって，物体，場所を指している場合が多い．一般に，指示の状態では，手を下ろした状態 (腰付近) から，離れた位置で指示を行うことが多い．

言語からの手がかりと同様，指示動作との共起性 Q_i を統計的に求め，以下で述べる指示動作抽出の際のスコアとして用いる．

4.3 注目要素の抽出処理

発話から指示詞，動詞を抽出し，上記スコアを足しあわせる．同様に，位置・動きデータからも上記特徴を抽出し，同様にスコアを足しあわせる．これらを簡単に表すと，以下のようになる．

$$S(t) = \sum S_i(t)$$

ただし，t は時刻，S_i は特徴 i に対するスコアである．それぞれの特徴に対するスコアは，前述の共起性 Q_i を基に決める．

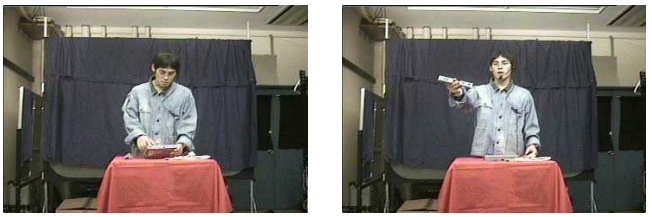
総合スコアがあるしきい値以上になる区間で指示，例示動作が起こっていると考える．次に，選ばれた区間から最もスコアの高い動作の静止点を現場指示動作の時点とする．その部分が注目フレームとなる．また，静止点がない場合には例示動作と考え，その区間全体を注目フレームとする．

予備調査として，既にデータベースに蓄えたプレゼンテーションに対して，以上の手法を適用した．その結果，位置・動きデータだけを用いて指示動作を検出した場合には，正答率は $30 \sim 40\%$，抽出率は最高で 70%弱となる．発話データのみを用いた場合[3]，位置・動きデータのみを用いた場合とは対照的に，80%程度の正答率が得られるが，抽出率は最大 60%程度にとどまる．これは，指示詞を伴わない指示動作があるためである．それに対し，発話と動作のスコアを単純に足し合わせた場合には，正答率は 80%弱となり，抽出率は 90%以上となる．発話のみを用いる場合に比べて，少し正答率が悪くなるが，抽出率は大幅に改善される．

4.4 実験例

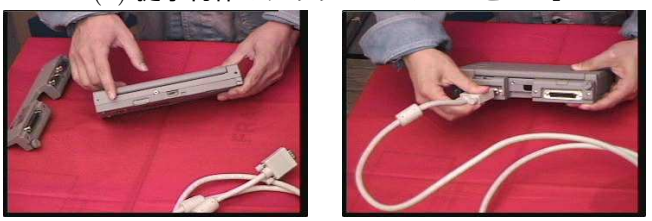
得られた注目要素を強調して提示することにより，注目を共有するための簡単なコミュニケーション補助システムができる．その例を図 12 に示す．

[3]音声認識結果ではなく，人手で書き起こした発話文を用いた



(a) 正面からの全体映像

(b) 提示物体のアップ 「この○ ○ を … 」



(c) 作業状態の提示 「このように○ ○ を△ △ します … 」

図 12: 提示映像の選択

(a) には正面からの映像が示されている．(b) では「この」と言いながら物体を提示する動作が行われたので，手元を重点的に提示している．(c) では「このように」と言いながら作業を行っているため，作業領域 (ここでは両手の間) を重点的に提示している．

このように，注目要素を選択的に提示することによって，相手の注意を有効にひきつけることができることがわかる．

現在の段階では，指示動作が行われてからその認識出力が得られるまでに，ある程度の時間がかかる．具体的には，

- 動作認識 $1 \sim 2$ 秒
- 音声認識 $1 \sim 3$ 秒
- 総合評価プロセス $1 \sim 2$ 秒

程度の時間がかかり，最終的に，$2 \sim 3$ 秒程度の遅れが生じる．動作認識の遅れ時間は，人間の動作に本質的に雑音が含まれ，指示動作であるとはっきりと判定できるまでに，少しの観察時間が必要であることが原因である．これらを高速化することが現在の課題となっている．また，遅れ時間が注意の共有に及ぼす影響について，調査を必要としている．

5　状況把握のための視覚支援

複合コミュニティ空間における自分や他人の状況を把握するために，仮想視点からの見え方を正確に生成する必要がある．我々は，複合コミュニティ空間中に複数台のカメラを設置し，それらによって撮影された多視点映像を用いて 3 次元物体の形状を推定し，仮想視点からの見え方を再現する研究に取り組んでいる．

5.1　3 次元形状推定

本研究では，3 次元形状を図 13 に示すような複数の平面上の形状，つまり 2 次元形状の集合として表現する．この手法では，平面間の線形補間によって 3 次元形状の解像度を向上させることが可能であるため，ボクセルモデルを中間媒体にする手法に比べ，高解像度の場合でも処理の高速化が容易であるという優位性がある．

次に，3 次元物体の形状を推定する手法について説明する．この時，カメラ位置 C_n において撮影された入力画像上の点 $I_n(u,v)$ と 3 次元空間中の平面 S_m 上の点 $S_m(i,j)$ の間には，以下の式で示すような 2 次元射影変換行列 P_n による関係が成り立っているものとする $(n = 1, ..., N, m = 1, ..., M)$ ．

$$I_n(u,v) = P_n S_m(i,j)$$

例えば，図 14 に示すように，水平 (XZ) 面に平行で高さ Y_1 の位置に平面 S_1 を設定し，そこに入力画像を 2 次元射影変換する場合，撮影した 3 次元空間中で実際に高さ Y_1 に存在する領域だけが平面 S_1 上に正確に射影される．この性質を利用して，平面上での物体の形状を推定する．まず全ての入力画像を平面 S_1 上に 2 次元射影変換し，それらの写像が重なる領域を求める．背景領域を除去した前景画像を射影する場合には，平面上で写像の論理積 (AND) をとることで物体領域を求めることができる．このようにして求められた領域の形状を 3 次元物体を平面 S_1 でスライスした形状であると仮定する．同様の処理を平面の位置，向きを様々に変化させながらくり返すことで，2 次元形状の集合として 3 次元形状を推定する．

5.2　自由視点映像の生成

獲得した 3 次元形状を用いた自由視点映像の生成手法について述べる．図 15 に，3 次元シーン中の注目点 $M(X,Y,Z)$ と各入力画像 I_n，自由視点画像 I_v との幾何的関係を示す．3 次元形状情報が既知であれば，隠れや表面の方向を考慮して注目点が観測された入力画像を割り出すことが可能である．また，自由視点映像を撮影するための仮想カメラの 3 次元位置，姿勢から，3 次元シーンと自由視点画像 I_v の関係を表す射影変換行列 P_v が算出できる．自由視点画像 I_v 上での注目点 $M(X,Y,Z)$ の見え方の再現法について説明する．まず，射影変換行列 P_v を用いて注目点の自由視点画像上での位置を求める．次に，3 次元形状情報を用いて各入力画像上における注目点の見え方を求める．それらの中から，3 次元物体による隠れの影響や物体の表面方向を考慮に入れ，自由視点からの見え方として最も適していると思われるものを選択し，自由視点映像上にマッピングする．これにより実際の観測画像に近い自然な見え方を再現することが可能になる．

5.3　実験結果

図 16 に本研究で構築した多視点映像撮影システムの構成を示す．今回の実験では，複合コミュニティ空間内にプログレッシブスキャン 3CCD カメラを壁・天井に合計 9 台設置した．これらのカメラには同期信号発生器から共通の外部同期信号が供給されている．このことにより，各フレームにおいてインターレースに伴う動物体のブレの発生がなく，かつ完全に同期したディジタル多視点映像の撮影が可能になった．更に，LTC(longitudinal time code) 発生器によって生成されたタイムコードをタイムスタンパ (TS) で VITC(vertical interval time code) に変換した後，各カメラのビデオ信号の垂直ブランク領域に書き込むことで，全てのカメラで撮影された個々のフレームの撮影時刻の特定が可能になっている．タイムコードを書き込んだ映像信号は，ビデオ取り込み用の画像ボードにより，ＰＣのメモリ上にディジタルカラー画像としてフルサイズ (640 × 480 画素)，ビデオレート (毎秒 30 フレーム) で取り込まれる．この撮影システムで撮影した映像の一部を図 17 に示す．

前節で述べた手法により生成される仮想視点からの見え方を図 18 に示す．この例では，仮想カメラを図左下に示すように移動させた．また，3 次元形状を表現するために 220 枚の平面を水平 (XZ) 面と平行に 1cm 間隔で配置している．

6　情報提示のための視覚支援

複合コミュニティ空間で行動する人間に対し，適所で適切な情報を提示する．そのために，ユーザ視点位置から撮影された実画像から，視点位置・姿勢を実時間で推定する手法の開発を行っている．本章では視点位置から撮影された実環境画像から視点位置姿勢情報を実時間で推定する手法について述べる．

本手法をユーザの視点位置画像を積極的に利用するタイ

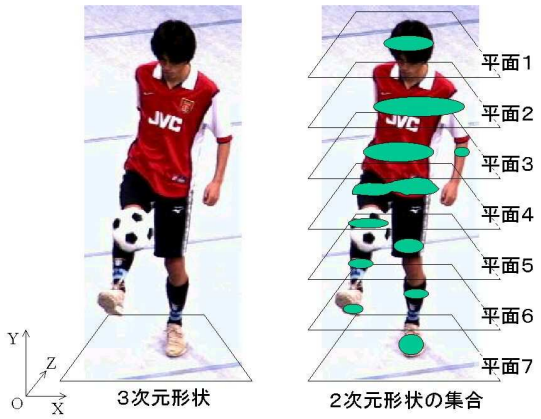


図 13: 平面による 3 次元形状表現

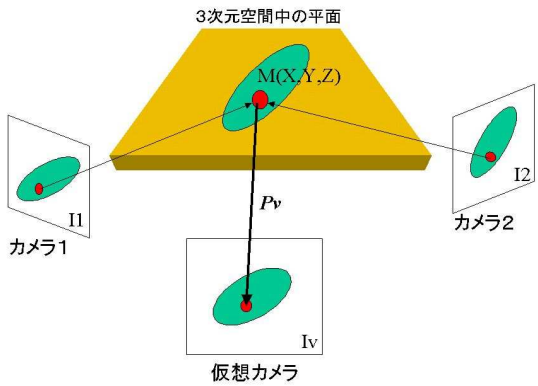


図 15: 自由視点画像の生成

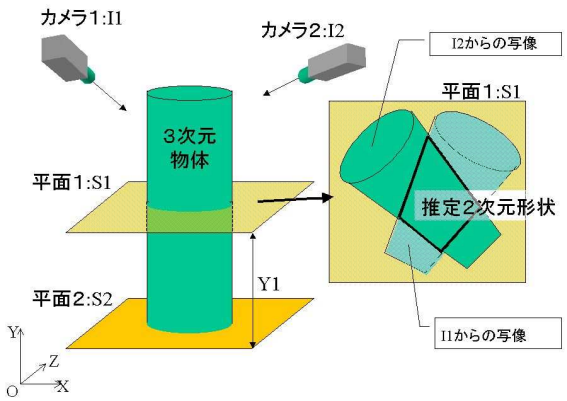


図 14: 3 次元形状推定手法

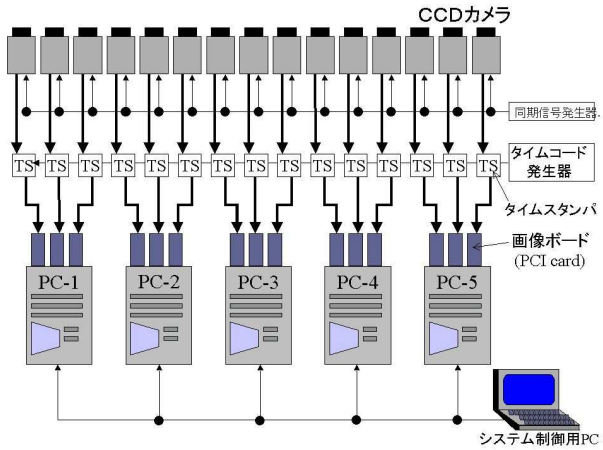


図 16: 多視点映像撮影システムの構成

プのウェアラブルコンピュータであるウェアラブルビジョンシステム [15] に適用することで，ユーザの着用機器を増やすことなくビデオシースルー型の複合現実感を提示することができる．従来，人工的なマーカを自然環境に配置したものを既知の特徴点として利用することで特徴点の追跡を簡略化していた [16]．しかしながら，マーカを特徴点として利用する限り，測定可能範囲を拡大するためにはマーカの設置範囲を拡大しなければならない．マーカを用いずに現実環境に存在する自然特徴点を利用することができるようになれば，特徴点の配置や認識方法を工夫することで測定可能範囲を自由に設定できるようになる． そこで本研究では人工的なマーカを用いずに現実環境に存在する自然特徴点をユーザの 3 次元視点位置姿勢の推定に用いるために，色ヒストグラムマッチング [17][18] とテンプレートマッチングを併用する自然特徴点探索手法を開発した．また，入力画像中で追跡に成功した特徴点の個数に応じて，撮影されたマーカの画面座標値からカメラの位置姿勢を求める複数の計算方法を切替えて用いる手法 [16] を用い，実際に 3 次元情報を合成する実験システムを試作した．

6.1 処理の概要

入力画像からのカメラ位置姿勢情報の取得処理は以下を繰り返すことで進行する．1. 入力画像から 3 次元位置が既知の複数の自然特徴点を探索・追跡し，画面座標値を求める．2. 抽出された特徴点の個数に応じた PnP 問題処理ルーチンでカメラの位置姿勢を推定する．3. 推定したカメラの位置姿勢をもとに，撮影されていない特徴点の画面座標を推定する．以降では，各段階での処理について述べる．

6.2 自然特徴点の追跡手法

カメラの位置姿勢を推定するために，入力画像から既知の特徴点を探索・追跡して画面上での 2 次元座標を求める必要がある．本手法では自然特徴点の情報を要素とするリスト (特徴点リスト) を事前に設定し，登録されている特徴点を既知の特徴点として入力画像中から探索・追跡して画面座標を求める．特徴点リストには事前に計測した 3

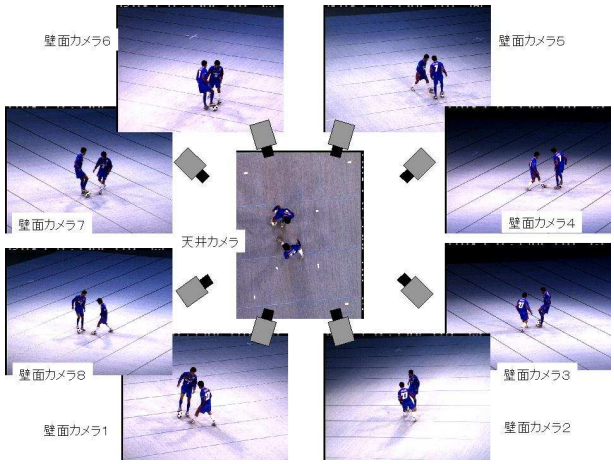


図 17: 撮影多視点映像

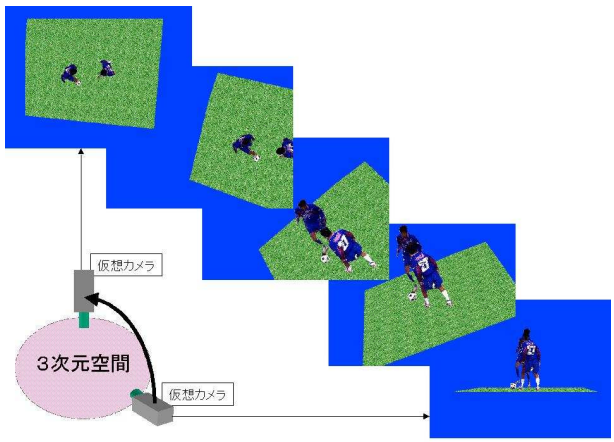


図 18: 生成された自由視点映像

次元座標情報と探索・追跡処理に用いられる画像特徴量が記録されている．

初期フレームでは入力画像全体を探索対象とするが，第 2 フレーム以降では，連続するフレーム間での特徴点の画面上の移動距離は小さいと仮定をおき，前のフレームでの画面座標を中心とする探索窓を設定する．

本手法では，精度を保ちつつ計算量を削減するために対応点探索を 2 段階に分けて処理する．

第一段階では入力画像の縮小画像を用い，高速処理が可能な色ヒストグラムマッチングにより対応点候補を絞る．次に第二段階では元のサイズの画像を用い，テンプレートマッチングによって輝度値の位置情報を用いた正確な対応点探索を行う．さらに，計算量の削減と安定した追跡を行うために第一段階でのマッチングに利用する色情報として (1) セグメンテーションに有効とされる I1I2I3 表色系 [19] を用いて輝度情報を分離し，色の情報 (I2,I3 値) による 2 次元ヒストグラムを，(2) 特徴点中心領域と周辺領域の 2 領域に分けてそれぞれに対して生成したヒストグラム (2 重色ヒストグラム) を利用する (図 19 参照)．これによ

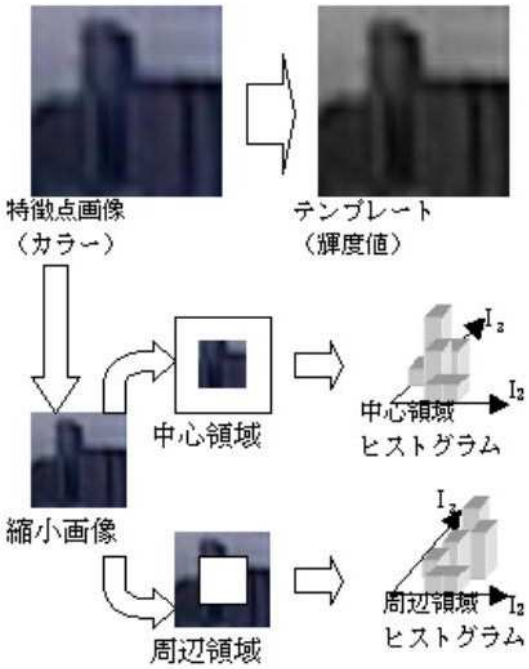


図 19: 探索に用いる画像特徴量の計算

り，通常の色ヒストグラムでは利用できない，色の位置情報をある程度保存したマッチング処理を行うことができる．ここで探索窓内に探索対象となる特徴点が見つからなかったと判定された場合 (マッチングの結果，類似度の最大値が閾値以下の場合) には，現実物体に隠蔽された，もしくは撮像範囲内から外れたと判断する．見つからなかった特徴点の画面座標は，特徴点の 3 次元座標が既知であるという仮定をおいているので，以降で求めるカメラの位置姿勢情報を用いて推定できる．この情報は次フレーム以降の追跡に用いられる．

6.3 カメラの位置姿勢の計算

カメラの位置姿勢は実世界に設定されている基準座標系からカメラ座標系への変換行列 (モデルビュー行列) で表すことができる．すなわち，ある点の基準座標系での座標値を，カメラ座標系での座標値を変換する行列である．本稿では前節で画面座標を獲得した特徴点の個数に応じて任意の 6 点，平面上の 4 点，任意の 3 点の画面座標値からカメラの位置姿勢を推定する手法 [20] [21][22] を選択して使用する手法 [16] を適用した．図 20 に各フレームでの処理の概要を示す．

6.4 観測されない特徴点の位置の推定

視点・視線の自由な変更を可能にし，カメラの位置と姿勢を安定に推定するためには，広範囲に存在する既知の

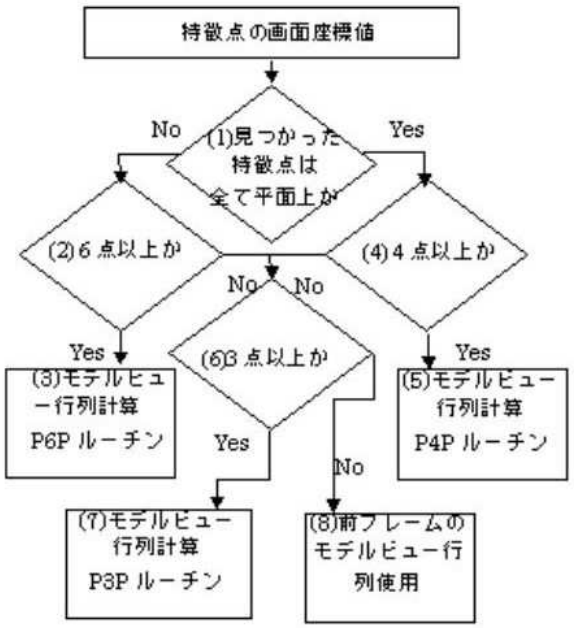


図 20: モデルビュー行列計算処理の概要

特徴点を順次切替えて利用する必要がある．このため，各フレームで抽出・追跡に失敗した特徴点についてもモデルビュー行列を用いて座標値を推定する．この処理により，実物体によって隠蔽されたり，撮像範囲外に存在するため追跡できない特徴点の位置を推定しておき，再び見えるようになったときや，撮像範囲内に新しく入って来た場合に追跡を再開もしくは開始する．

6.5 実験結果

前節で提案した実時間カメラ位置姿勢推定手法の動作を確認するために，ウェアラブルビジョンシステムのヘッドセットと動画像の入出力および動画像のハードウェアテクスチャマッピング機能を持つ PC(SGI VWS540 PentiumIII Xeon 500MHz 4CPU) を用いて動作確認実験を行った．提案手法を用いて 3 次元 CG を合成した画像提示の様子を図 21 に示す．この例ではコアラのぬいぐるみの CG モデルを，その CG 作成手法の研究紹介用のパネルに対する付加情報として観察者に提示している．いわば飛び出すポスターである．合成環境を提示する手法にはビデオシースルー方式による合成画像提示を採用した．試作システムの画像更新レートは，実行例に示す程度の仮想物体 (約 4000 ポリゴン) を合成する場合で約 5 ~ 10 フレーム / 秒である．入力画像の解像度は 360 × 243 画素，このとき追跡していた特徴点数は 4 点で，特徴点領域，探索窓領域はそれぞれ 24 × 24 画素，48 × 48 画素であった．

今後は，追跡特徴点の自動選択，特徴点検出開始の自動化，推定精度の向上・高速化の研究を行い，実世界シーンを実時間で扱うウェアラブル型の視覚情報メディアシステムを開発する予定である．

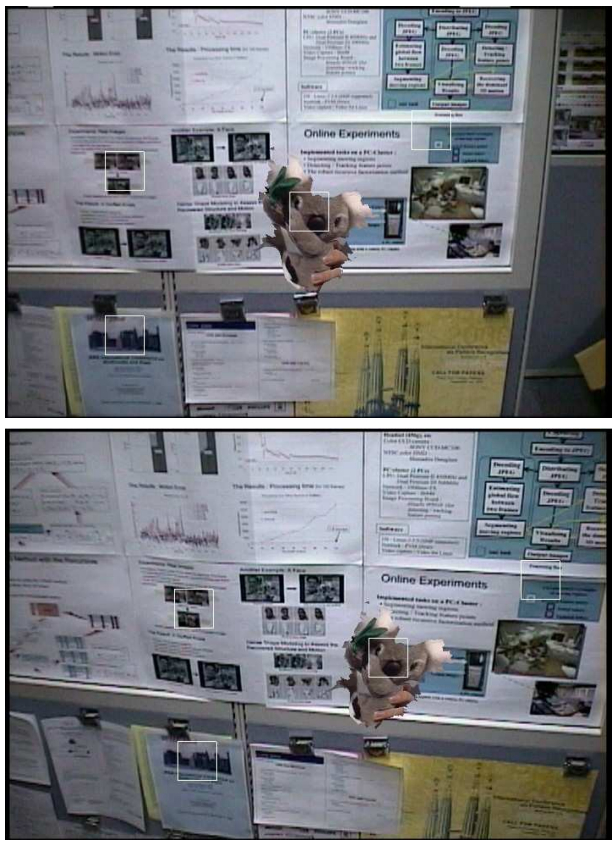


図 21: 研究紹介用ポスターに実験結果の 3 次元 CG モデルを付加 (飛び出すポスター)

7 まとめ

複合コミュニティ空間の枠組みとその要素研究について簡単に述べた．今後，個々の要素技術を完成させていくとともに，実際に複合コミュニティ空間として動作可能なプロトタイプシステムの構築を開始する予定である．

顔映像の加工については，アイコンタクトに特に重要な眼の表情を再現した顔映像を，HMD をかけた現実像へ重畳する技術を開発する．眼の動きは，我々が新しく製作した HMD 内蔵型の視点検出装置を用いて計測するが，微妙な眼の表情を再現し任意の向きの顔に実時間で重畳するためには，リアルタイム性を考慮した高度のパターン認識・理解技術が必要であると考えられる．

注目の共有については，種々の動作が行われた場合に，重要なパラメータ (例えば，指示方向，大きさ) を抽出する方法，それを基に提示情報を制御する方法について研究する．具体的には，手指の方向，形，視線などから，注目すべき位置，方向，対象，大きさ等の情報を抽出し，複合現実空間中の物体や場所を特定する手法を開発する．指示動作のような明示的な動作だけでなく，視線方向のような無意識の動作から得られる情報を統合し利用することが重

要であると考えられる．

参考文献

[1] 横矢直和, 大田友一: “視覚増強のためのパターン認識・理解”, MIRU2000 論文集, 2000

[2] 向川康博, 中村裕一, 大田友一: “複数の顔画像の組み合わせによる任意方向・任意表情の顔画像の生成”，信学論，Vol.J80-D-II, No.6, 1997.

[3] 大和哲也, 大田友一: “実写画像に基づく顔映像の生成と実映像との合成”, MIRU2000 論文集，2000

[4] 大和哲也, 大田友一: ”実写画像に基づく顔映像の生成”,2000 年電子情報通信学会総合大会,(2000)

[5] 大田友一: “画像を手がかりとした実世界と仮想世界の幾何的整合”, 信学技報,PRMU98,251, (1999)

[6] 中村裕一, 西谷正志, 大田友一: “プレゼンテーション映像における話者の行動理解” 信学技報 PRU95-143, 1995

[7] Y.Nakamura, Y.Kimura, Y.Ye, Y.Ohta: “MMID: Multimodal Multi-view Integrated Database for Human Behavior Understanding” Int . Conf . on Automatic Face and Gesture Recognition, 1998

[8] Y.Nakamura: “Multimodal Approach toward Intelligent Video Production”,Proceedings Multimodal Approach toward Intelligent Video Production,(1999)

[9] 中村裕一: “コミュニケーションのための画像・映像処理”, 信学技報, PRMU99,252,(2000)

[10] P. Ekman and W. Friesen: “The Repertoire of Nonverbal Behavior : Categories, Origins,Usage,and Coding”, Semiotica, 1, 1969

[11] 北原格, 大田友一, 斎藤英雄, 秋道慎志, 尾野徹, 金出武雄: “大規模空間における多視点映像の撮影と自由視点映像生成”,3 次元画像コンファレンス 2000,(2000)

[12] 北原格, 大田友一, 金出武雄: “多視点映像の融合によるスポーツシーンの自由視点映像生成”, MIRU2000 論文集, 2000

[13] 北原格, 大田友一: ”多視点映像の融合による運動する物体の自由視点映像の生成”,2000 年電子情報通信学会総合大会,(2000)

[14] 藤村祐介, 三角田健作, 大田友一: ”全天周カメラと注視点制御カメラを用いた動物体を含む 3 次元環境モデルの獲得”,2000 年電子情報通信学会総合大会,(2000)

[15] 蔵田, 興梠, 栗田, 村岡, 坂上: “気の効いたウェアラブルビジョンシステムの実現に向けて：顔検出・識別とハンドマウスの実装”, 信学技報 Vol.99, No.710, PRMU99-270, 2000

[16] 大隈, 竹村, 横矢: “拡張現実感システムのための画像からの実時間カメラ位置姿勢推定”, 信学論 D- II, Vol.J82-DII, No.10, 1999

[17] M. J. Swain, and D. H. Ballard: “Color Indexing”, International Journal of Computer Vision, Vol.7, No.1, 1999

[18] 村瀬, and V. V. Vinod: “局所色情報を用いた高速物体探索–アクティブ探索法–” 信学論 D-II Vol.J81-D-II No.9, 1998

[19] Y. Ohta, T. Kanade, T. Sakai: “Color Information for Region Segmentation,” Computer Graphics and Image Processing, Vol.13, 1980.

[20] 出口: “射影幾何学による PnP カメラ補正問題の統一的解法”, コンピュータビジョン’90 シンポジウム, 1990.

[21] 中沢, 中野, 小松, 斎藤: “画像中の特徴点に基づく実写画像と CG 画像との動画像合成システム”, 映像情報メディア学会誌, Vol.51, No.7,1997.

[22] R. M. Haralick, C.-N. Lee and K. Ottenberg: “Analysis and solutions of the three point perspective pose estimation problem,” Proc. CVPR ‘91, 1991.