日本バーチャルリアリティ学会第 6 回大会論文集 (2001 年 9 月)

複合コミュニティ空間における注目の共有 ～人物動作理解による物体への注釈付け～

Sharing Attention in Mixed Community Space
— Recording Annotation for Objects by Recognizing Human Behaviors

尾関基行 [1)]，伊藤雅嗣 [1)]，中村裕一 [1)2)]，大田友一 [1)]
Motoyuki OZEKI, Masatsugu ITO, Yuichi NAKAMURA and Yuichi OHTA

1) 筑波大学 機能工学系
（〒 305-8573 つくば市天王台 1-1-1, ozeki@image.esys.tsukuba.ac.jp）
2) 科学技術振興事業団 さきがけ研究 21

Abstract : *The Mixed Community Space* is a space where multiple people share a mixed reality world and communicate each other. We handle not only virtual objects but also a virtualized real world, and share important information even over different spaces or time. In this paper, we propose a new method for sharing attentions among people by detecting and capturing typical situation which a person is giving important explanation on an object, and by giving video record to other people. This paper presents the basic idea, the automated cameras simultaneously tracking important portions, and the annotation capturing mechanism.

Key Words: *Mixed Reality, Mixed Community Space, Camerawork, Human Behavior Recognition, Multimedia Contents*

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ことのできる人工空間である．この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に加え，現実世界にはない新しい視覚情報を同時に共有する．本稿では，このような空間における人間同士のコミュニケーションを補助するための機能について述べる．

一般に複合現実空間システムでは，HMD 等のデバイスを装着することや仮想物体やその他の情報が重畳されて視覚提示されることにより，現実空間に比べてノンバーバルなコミュニケーションが難しくなるという問題がある．例えば，目が HMD によって隠されるためにアイコンタクトが妨害され，自分の意図が相手に伝わらないことがある[1]．また，あるユーザが物を指し示して説明していることに，他のユーザが気付かないといったことが起こり得る．

この問題に対し，本研究では，複合コミュニティ空間における「注目の共有」を支援するための機能を実現する手法を提案する．例えば，ある人が注目しているものを他の人に強調して提示することによって，話題の中心を明確にし，意志の疎通を円滑にすることを目的としている．さらに，注目が必要とされている場面を映像として記録し，それを物体と関係付けることによって，時間や場所が異なる場合にも注目対象に関する情報が利用できるような枠組みを提案する．

以下，本稿では，これらの機能についての考え方とプロトタイプシステムの構成について説明し，その後に注釈情報の記録について詳しく述べていく．

[1]我々のグループでは，これに対し，HMD によって隠される人物のアイコンタクトを復元する研究を行っている [1]

2. 複合コミュニティ空間のための注目共有

2.1 注目共有の考え方

本研究では，複数人が複合空間の中で対話している場面，特に，その中の一人が物やその利用方法について説明している場面を想定する．このような場面において，注目が必要となる状態を検出し，撮影した映像を他のユーザに強調提示したり，注釈情報として記録しておく．このようなことを実現するために必要となる機能は，次のようになる．

(a) 注目対象 (候補) を複数カメラで追跡・撮影する機能:
注目対象となり得る部分，例えば，人物の主要部位，主要物体，場所等を自動的に追跡することが必要である．また，人間に提示する映像を撮影することから，プレゼンテーションや説明シーンを撮影するために適切なカメラワークも必要となる．

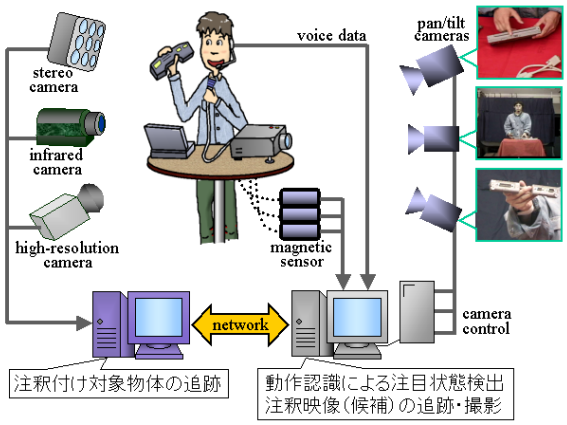

図 1: 物体への注釈付けシステム

(b) 注目が必要となる状態を検出する機能: 人物 (話者) の発話や動作を認識して，注目が必要とされている状況とその時の注目対象を認識する必要がある．どのような動作を検出する (利用する) かは目的にもよるが，本研究では，まず，人物が意図的にはっきりと注目を要求する場面を対象とする[2].

(c) 撮影映像を注目対象と関係付けて記録する機能: 注目対象に与えられた説明やその際の人物の行動を撮影し，それを物体 (位置やテクスチャ) と関係付ける．注釈情報としては，注目対象の名前，位置，使い方やそれに関する過去の出来事などが重要となる．同じ時刻に行動している他のユーザに対しては，撮影されている映像を実時間で提示するだけでもコミュニケーションの補助となるが，時間を隔てた場合のために，映像クリップとして一旦蓄積する．

(d) 注目対象の強調提示や注釈情報としての提示機能: 同時刻に行動している他のユーザには，(c) で得られる映像をリアルタイムで提示できる．また，時間を隔てた場合にも，他のユーザが当該物体への興味を示した場合に，蓄積しておいた過去の説明映像を提示することができる．(a) と同様，物体 (注目対象) の位置を常に検出しておくことが必要となる (これについては文献 [2] で詳しく論じる).

具体的な例として，ある人物が物体を指差しながら，それの利用方法を説明する場合を考えてみよう．まず，複数のカメラがその人物がいる空間，手先，その他の重要部分を，各々の対象に適したカメラワークで追跡・撮影している．ある時刻にその人物が物体を指示しながら，その物体に関する説明を始めると，システムはその動作を認識し，注目が必要な状態として検出する．これをトリガとして，各カメラにより撮影されている映像からその説明を最も良く捉えている映像が選択され，指示された物体と関連付けて記録される．これらの映像は，同時刻に複合コミュニティ空間内にいる他のユーザには実時間で提示され，また，それ以外のユーザのためには物体の位置，テクスチャと関係付けられて蓄積される．

[2]例えば，物体を掲げて「この○○は…」と発言するような動作

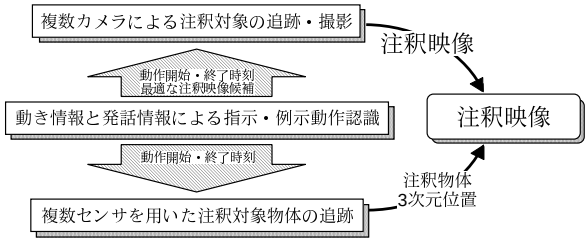

図 2: システムの概要

表 1: “対象物” と “対象の状態”

(a) 対象物	
人物自身	人物の全体的な様子を狙う
作業空間	作業が行われている空間を狙う
注目物体	注目すべき (注釈付けされる) 物体を狙う
注目場所	注目すべき (注釈付けされる) 場所を狙う
(b) 対象の状態	
＜状況＞	シーン中での位置関係や軌跡
＜操作＞	操作等, 細かい動きが行われている状態
＜物体＞	対象物自体の状態 (形, 色, 静止状態等)

2.2 システム構成

前節で述べたことを実現するために，図 1 に示すようなプロトタイプシステムを構築した．まず，人間の動作を計測するために，接触型の位置センサ (磁気センサ) を用いている．これによって人体の各部位の位置を計測し，複数台の首振りカメラで主要部分を追跡・撮影する．また，複数の画像センサ (可視光カメラ，ステレオカメラ，赤外線カメラ) を併用して，話題の対象となっている物体を追跡する．これにより，物体の 3 次元位置を常に取得しておき，撮影システムにより得られた注釈映像と関連付けることを可能にする．

発話情報は音声認識エンジンを用いて取得する．現在は連続音声認識の結果から簡単なキーワードを抽出して用いている．動作認識の部分では，人体の各部位の位置情報より計算される動き情報と，音声認識ソフトによって得られる発話情報を統合することで，物体への指示・例示動作を検出する．この処理の全体的な流れを図 2 に示す.

2.3 追跡・撮影のためのカメラ制御

注釈映像の撮影では，注目すべき対象を適切な大きさ・位置で画面に捉えることに加え，人間にわかりやすく情報を伝えるためのカメラワークが必要となる．つまり，人物や物体をただ単純に追尾撮影すれば良いわけではなく，撮影する対象やそこで起きているイベントに応じたカメラ制御を行う必要がある．そのために，我々は撮影対象を「何のどういう状態に注目するか」，つまり，撮影したい “対象物” と撮影したい “対象の状態” に分けて考えることで，撮影対象とカメラ設定の関係を簡潔にまとめた．

“対象物” と “対象の状態” をそれぞれ表 1(a), (b) に示す．撮影の際に，対象物を (a) の 4 つから選択し，さらにその撮影範囲を “大・中・小” の中から選択することにより，追跡する対象基準点と解像度が決定される．また，選択した対象物のどういう状態に注目するかを (b) の 3 つから選択するこ

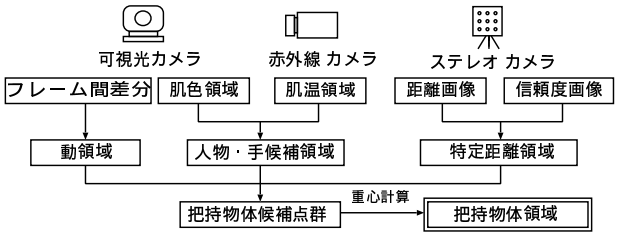

図 3: 把持物体検出の流れ

とで，それぞれの状態に応じたカメラ制御方法で追跡撮影を行う．

本研究では，対象に応じた適切なのカメラワークを実現するために，**カルマンフィルタによる平滑化**と**枠制御アルゴリズム**を用いている．これらの中で定義されたいくつかのパラメータを調節することで，目的に応じたカメラワークを実現することができる．また，ノイズによる細かいブレや，手などを撮影対象とした場合の細かな動きによるカメラ振動といった問題も解決している．

紙面の都合上，簡単に概要を説明したが，詳細については文献 [3] を参照されたい．

3. 把持物体の認識と注釈情報の記録

3.1 複数センサを用いた把持物体の追跡

2.1 で述べた機能を実現するためには，注釈付けの対象となる可能性のある物体を常に追跡しておくことが必要となる．また，複合コミュニティ空間に種々の物体を持ち込むことを可能にするため，**物体の大きさ，色等に関する予備知識はない**という前提の下での追跡が好ましい．しかし，このような条件で任意の物体を認識することは一般的に難しいため，本研究ではまず，人物が手に持っている物体 (把持物体) に限定し，それ以外の場合は今後の課題とする．

このような場合の物体検出・追跡方法として，本研究では，異なる種類の画像センサを相互補間的に用いる．把持物体抽出の考え方は次のようになる．まず，本研究では，手と近接していて，手と同じ動きをする手以外の部分を把持物体と考える．つまり，把持物体領域を抽出するためには，手を抽出し，これと同じ動きをする手以外の部分を抽出すれば良い．この目的のために，本研究では，可視光 (RGB) カメラ，赤外線カメラ，ステレオカメラを用いる．これらの画像センサから得られた画像より，肌色領域，肌温領域 (赤外線カメラによって検出される温度の比較的高い領域)，特定距離領域 (手と把持物体が存在する可能性の高い距離にある領域)，動物体領域を抽出し，これらを組み合わせて手と把持物体領域の抽出を行う．手領域は「肌色領域 ∧ 肌温領域 ∧ 特定距離領域 ∧ 動領域」とし，把持物体領域は「特定距離領域 ∧ 動領域 ∧∼ 手領域」とする．

把持物体検出の流れを図 3 に示す．ステレオカメラからの距離画像を用いることにより，複合コミュニティ空間に複数の人が存在し，背景中で他の人物が動いている場合や物体が動いている場合でも対処できる．また，赤外線画像を用いることにより，環境中に肌色に近い領域が存在しても，誤検出することが少なくなる．

図 4: 各処理の結果と把持物体の追跡結果

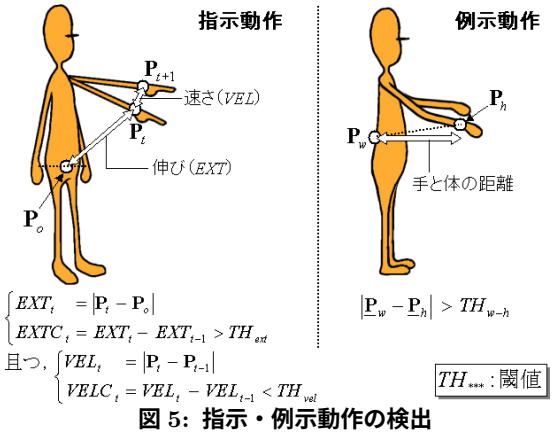

図 5: 指示・例示動作の検出

以上の方法で実験を行った例として，それぞれの画像センサから得られる画像と追跡結果を図 4 に示す．上段左側が可視光カメラから得られた肌色領域，右側が赤外線カメラから得られた画像，下段左側がステレオカメラから得られた距離情報である．そして，下段右側が追跡中の物体領域に枠を上書き表示したものである．

3.2 動き情報と発話情報による指示・例示動作認識

人物が意図的に注目を誘導する場合には，指示・例示動作がよく現れることが知られている．本研究では，指示・例示動作の検出を発話と動作の両方を処理することで行い，注釈付けのトリガとして用いる．

発話情報としては，まず，指示・例示動作を行う際によく現れる指示詞を利用する．これに人物の動き情報を併用することで，指示詞からだけでは得られない “動作の行われた場所” を知ることが可能となり，また，文脈指示詞などによる動作の誤検出を防ぐことができる．

図 5 に動き情報の検出方法を示す．指示・例示を行う動きとは，手を大きく振り上げて急停止する動きであると定義し，手を降ろした位置からの伸びの変化と手の加速度を用いて検出している．表 2 に認識対象としている動作について，各々対応する指示詞と話し手の動きをまとめたものを示す．詳細については，文献 [4] を参照されたい．

表 2: 動作の分類と対応する発話・動き

動作の種類	発話の種類	併用する動き
例示動作	このように こうやって, 等	手が体から離れている
物体指示動作	これ系統 この＋名詞	指示・提示を行う動き
場所指示動作	ここ系統 この＋場所	指示を行う動き

4. 実験例

以上の要素技術を組み合わせて，実際に物体への注釈記録した例を示す．内容は一人の人物が机の上にある料理を順に説明していくことを模擬したものである．

得られた映像を図 6 に示す．図は把持物体追跡の結果を示しており，左下に注釈映像の候補として選択されている映像を表示している．「把持物体の検出→動作の検出→注釈映像との関連付け」という流れに対応して，把持物体上に表示されている枠が「破線→太い赤実線→実線」と変化していく．また，注釈情報も動作認識をトリガとして，適宜最適と思われる映像が選択されている．

注釈映像としては，物体が持ち上げられた時点から，説明が終了し，再び物体が置かれたところまでを 1 クリップとした．結果より，意図した通りに物体への注釈付けが行われていることがわかる．

以上の実験では，ある程度大きな物体に対する比較的検出しやすい動作が行われているため，上記のような結果が得られているが，任意の物体に対して自然に振る舞った場合にシステムがうまく動作するためには，まだ多くの点で改良が必要である．今後，種々の面から実験を重ね，改良していく予定である．

5. まとめ

複合コミュニティ空間における注目の共有のために，物体への注釈付けを行う枠組みの提案を行った．この中で，複数カメラによる対象の追跡・撮影，人物動作認識による注目状態の検出，複数センサによる把持物体の追跡という 3 つの課題について取り組み，プロトタイプシステムによる実験を通してそれらの手法の有効性を確認した．

今後は，システムの不安定な部分を解決していくと共に，注釈情報として映像と位置以外にも様々なものを付加していくことを考えている．また，遠隔地のユーザも交えた実時間での注目共有や，実際に HMD を装着した複合コミュニティ空間での共同作業など，様々な場面へ適用していく予定である．

(a) 図は把持物体の追跡経過を表示した例．左下に上書きされている画像は，実時間で相手に提示される映像．

(b) 把持物体が検出されたところから注釈情報が記録される．検出された把持物体 (下中央) の上に枠が表示されている．

(c) (a) の発話情報と (b) の動き情報から指示・例示動作が検出されると，それに関する情報 (開始・終了時刻，動作の種類) が把持物体と関連付けて保存される．

(d) 把持物体が机に置かれたところで注釈情報の記録を終了し，把持物体の位置とテクスチャが保存される．

(e) 前に検出された物体の情報は保持したまま，新たな把持物体が検出され，同様の処理が始まる．

図 6: 物体への注釈付けの結果

参考文献

[1] 竹村 雅幸, 北原 格, 星野 准一, 大田 友一：複合コミュニティ空間における人物映像加工によるアイコンタクトの復元，VRSJ 第 6 回大会論文集，2001.

[2] 里 雄二, 北原 格, 中村 裕一, 大田 友一：複合コミュニティ空間における注目の共有 ～指示動作による注目の強調提示システム～，VRSJ 第 6 回大会論文集, 2001.

[3] 尾関 基行, 中村 裕一, 大田 友一：プレゼンテーションの知的撮影システム —手元作業を対象とした適応的カメラワーク—，PRMU2000-104，pp. 31-38，2000.

[4] 尾関 基行, 中村 裕一, 大田 友一：プレゼンテーションの知的撮影システム —動作認識による映像のタグ付け—，IIM2000(6th)，pp. 69-74，2000.