映像解析による人物動作理解に関する研究

髙橋　正樹

博士（情報学）

総合研究大学院大学
複合科学研究科
情報学専攻

平成23年度
（2011）

2012年3月

本論文は総合研究大学院大学複合科学研究科情報学専攻に
博士（情報学）授与の要件として提出した博士論文である.

審査委員：

佐藤　真一（主査）	国立情報学研究所／総合研究大学院大学
孟　　洋	国立情報学研究所／総合研究大学院大学
佐藤　いまり	国立情報学研究所／総合研究大学院大学
佐藤　洋一	東京大学
杉本　晃宏	国立情報学研究所／総合研究大学院大学

（主査以外はアルファベット順）

A study on human action understanding based on video analysis

Masaki Takahashi

DOCTOR OF
PHILOSOPHY

Department of Informatics
School of Multidisciplinary Sciences
The Graduate University for Advanced Studies (SOKENDAI)

March, 2012

A dissertation submitted to the Department of Informatics,
School of Multidisciplinary Sciences,
The Graduate University for Advanced Studies (SOKENDAI)
in partial ful¯llment of the requirements for
the degree of Doctor of Philosophy

Advisory Committee

Shin'ichi Satoh (Chair)	National Institute of Informatics / The Graduate University for Advanced Studies
Hiroshi Mo	National Institute of Informatics / The Graduate University for Advanced Studies
Imari Sato	National Institute of Informatics / The Graduate University for Advanced Studies
Yoichi Sato	The University of Tokyo
Akihiro Sugimoto	National Institute of Informatics / The Graduate University for Advanced Studies

(Alphabet order of last name except chair)

論文要旨

近年では，個人が所有するＰＣやＴＶにもカメラが搭載されるようになり，映像を用いたコミュニケーションが一般的となっている．またインターネット上には大量の映像が溢れ，様々な映像を誰もが視聴できる環境が整っている．街中にはいたるところに監視カメラが設置され，屋内外の映像を常時記録している．このように，現代では映像が様々な形で日常生活の中に浸透しており，簡単にアクセス可能な身近なメディアとなっている．

本論文はこのような実環境における映像を対象とした，人物動作認識技術の高度化を提案するものである．生活環境への映像の普及に伴い，ユーザ動作による機器操作，特定動作をクエリとした映像検索など，人物動作認識技術に対する期待は高まっており，その需要も多岐にわたる．しかし実環境における映像は撮影条件や被写体動作が多様であり，安定した解析が困難であることが多い．本研究はこれらの課題を確認するとともに，頑健な人物動作認識技術の確立を目指すものである．あわせて，人物動作から動作者の意図や内部状態までも理解することを目指した．

第1章では本研究の背景として，日常生活における人物動作認識への社会的ニーズについて述べる．また人物動作の多様性に対処するため，意図の強さに基づく人物動作の分類を行う．分類した各動作の認識技術に対するニーズを確認するとともに，その実現に向けた課題を検討する．最後に，映像解析による一般行動認識技術についてまとめ，関連技術の現状を確認する．

第2章では意思伝達動作であるジェスチャに焦点をあて，その認識手法について検討する．近い将来の大画面・高解像度のＴＶ視聴環境では，リモコンに代わる新たなマンマシンインタフェースが求められている．中でも映像解析による人物ジェスチャ認識は接触型デバイスが不要であり，次世代ＴＶの特徴である没入感を損なうことなく操作できるため，新たなインタフェースとして期待を集めている．またジェスチャによる操作は，映像コンテンツ内オブジェクトとの自然なインタラクションを実現するうえでも有効である．本章では，はじめに次世代ＴＶ視聴環境でのジェスチャ認識における要件を考察する．続いて，ジェスチャ認識における先行研究を紹介し，本章での研究目的を明確にする．具体的には，次世代ＴＶ視聴環境での対話型ジェスチャ認識の実現へ向け，奥行き情報の利用方法について検討する．またユーザの自然な動作の認識を実現するため，長期の時間情報を含む画像特徴を検討する．続いて，これらの検討を踏まえた新た

なジェスチャ認識手法を提案する．最後に，様々な実験を通して提案手法を評価し，次世代ＴＶ視聴環境でのインタフェースとしての有効性を確認するとともに，今後の拡張性を考察する．

第3章では人物の一般行動に焦点をあてる．人混みで混雑した実環境での監視映像を対象とし，混雑映像から一般行動を頑健に認識する手法を提案する．現代では屋内・屋外問わず監視カメラが普及しているが，その映像の確認はほとんどの場合人間によって行われている．膨大な映像量に対する作業者の数は少なく，非効率な監視を余儀なくされている．またそのほとんどは映像の事後確認にとどまり，犯罪の未然防止や直後の検出には活かされていない．そのため，不特定多数の人物行動を自動認識する技術への期待が高まっている． 本章ではまず混雑映像の解析における問題点を列挙し，その課題を確認する．続いて一般行動認識に関する先行研究を紹介し，関連技術の現状について述べる．そして広域特徴に基づく手法，局所特徴に基づく2つの手法について検討する．前者は，人物領域検出の結果に基づき人物の軌跡からその行動を認識する手法である．後者は，特徴点軌跡に基づき多数の軌跡特徴から人物行動を認識する手法である．2手法の比較を通し，実環境で有効に機能する人物行動認識手法を検討する．最後に，独自の様々な実験による評価と，映像検索に関する国際的評価型ワークショップTRECVID Surveillance Event Detectionタスクへの参加を通し，提案手法の有効性を確認するとともに，実用化へ向けた課題を確認する．

第4章では，ＴＶ視聴者の個人的趣味・嗜好を理解するため，映像視聴中のユーザの筋運動系情動から内部状態（注目度）を推定する手法を検討する．ユーザの内部状態は情動として一部身体に表出すると考えられるが，その動作は微小であり，正確に計測することは難しい．さらに計測した情動動作がそのまま内部状態を表しているとは限らない．たとえば映像コンテンツ内特徴など，外的要因も考慮する必要がある．これらの課題により，可視情報からの内部状態推定は一般に困難とされてきた．本章では，はじめに脳科学や心理学など，他分野研究を含めた人物の内部状態推定に関する先行研究を紹介する．次にユーザが注目状態にあるときに表出する情動動作を検討し，それら動作と注目度に関する仮説を示す．目視正解データによる仮説の検証の後，各情動特徴を自動取得する手法を提案する．最後に，自動計測した情動特徴から注目度を算出する注目度推定器を作成し，その性能を評価するとともに，提案手法の将来性を考察する．

第5章では，本論文の成果をまとめる．本論文の成果は，実環境における人物動作認識のニーズとその実現へ向けた課題を確認し，各ニーズにおける人物動作認識手法を提案するとともに，その有効性を検証したことにある． さらに意図により動作を分類し，各段階での知見を活用しながらより微小動作の認識手法を提案し，動作者の意図を理解する手段を示したことにある．特に無意図的動作と呼ばれる情動動作から人物の潜在的

興味を推定する可能性を示せたことは，大きな成果であると考える．従って本研究は，家庭における新たなマンマシンインタフェースや個人プロファイル推定，監視カメラをはじめとする各種映像における人物行動検出など，実生活環境で利用可能な人物動作理解技術の確立へ向け，大きな貢献をしたと考える．

Abstract

Videos have become in various ways a part of daily life and are now an easily accessible form of media. Many personally owned PCs and TVs are now equipped with a camera, leading to the wide use of videos in communication. The Internet is also teeming with vast amounts of video images that are widely accessible to anybody. In many cities, surveillance cameras are installed in different places to constantly record videos inside and outside buildings.

This paper demonstrates advances in technologies for recognition of human motion in videos taken in actual environments. With the wide use of videos in daily life situations, there has been an increasing demand for human motion recognition technologies, such as for operating devices through user movements and for doing video searches using specific movements. Videos from actual environments, however, are difficult to analyze reliably because of variations in shooting conditions and object movements. This study thus aimed to look at these issues and establish reliable human motion recognition technologies.

Chapter 1 gives a background of the research, mentioning society's needs for human motion recognition technologies in daily life. The chapter discusses the classification of human motions in an effort to address their diversity. Specific technologies needed for recognizing motions in each category are then discussed, and their feasibility is assessed. The chapter concludes with a discussion of general motion recognition technologies based on video analysis and other related technologies.

Chapter 2 focuses on gestures, which serve as a medium of communication, and discusses the methods for recognizing them. Recognizing gestures is important in the light of the demand for a man-machine interface that will replace remote controls in large-screen high-definition TV viewing in the near future. Since touch-type devices are not needed and operations can be made without losing immersive quality, which is a characteristic of next-generation TV, human gesture recognition technologies based on image analysis are getting wide attention as a new form of interface. Gesture-based operations also enable a means for natural interaction with objects within videos. The chapter begins with a discussion of the requirements for gesture recognition in next-generation TV viewing. Next, it introduces previous researches on gesture recognition and outlines the objectives of the research discussed in the chapter. In particular, the research aims to study depth-information usage methods that are necessary for realization of

interactive gesture recognition needed in next-generation TV viewing environments. It also aims to study image features that include long-term temporal information to enable a natural recognition of user's movements. A new gesture recognition method is then proposed from these studies. Likewise, the effectiveness of the proposed gesture recognition method in serving as an interface for next-generation TV viewing environments is ascertained and evaluated through various experiments, and a discussion of its scalability is made.

Chapter 3 focuses on general human behavior. The chapter proposes a reliable method for recognizing general behavior using actual videos of people in crowded areas taken through surveillance cameras. Although outdoor and indoor surveillance cameras have become widely used nowadays, image recognition is usually done manually. Since there are only few operators relative to the large amount of video data that need to be processed, surveillance could not be effectively done. In addition, since these videos are not processed immediately, they do not contribute to crime prevention and are not being used for crime detection at time of occurrence. Thus, there is an urgent need for a technology for automatic recognition of large numbers and types of human behavior. This chapter enumerates problems and discusses issues related to the analysis of images of people in crowded areas. It also mentions previous research on recognition of general human behavior and provides an overview of related technologies. It also discusses the two main types of methods for recognition of human behavior, namely, the global-feature-based method and the local-feature-based method. The former is based on detection of human regions, wherein human behavior is recognized on the basis of the trajectory of the human region. The latter is based on feature-point trajectories, wherein behaviors occurring within the video are recognized from multiple trajectory features. A human behavior recognition method that functions effectively in actual situations is proposed from these two methods. The chapter also includes a discussion of issues related to practical application of the proposed method as well as a discussion of its effectiveness as evaluated through various propriety experiments and through participation in the TRECVID Surveillance Event Detection evaluation task, an international evaluation workshop for event detection in video surveillance.

Chapter 4 discusses methods for estimating internal state (attentiveness) of the user viewing the video in accordance with image features obtained through measurements of the viewer's emotions expressed through certain muscular movements. Generally, the user's internal state is partly expressed through his or her body as emotional behavior. These actions, however, are usually imperceptible and are difficult to measure accurately. In addition, they are not necessarily straightforward expressions of the user's internal state, making it important to consider external factors such as content features of the video. As such, inferring internal states

on the basis of visual information is usually a daunting task and can be considered as a very challenging topic. The chapter begins with an introduction of previous researches on inference of human internal states, including those in the fields of neuroscience and psychology. Next, it discusses emotional behaviors that are observed when users are paying attention and mentions some hypotheses regarding the relationships of these behaviors to user attentiveness. After verification of these hypotheses using correct visual data, a method for automatically obtaining the different emotional features is proposed. Lastly, it introduces the fabrication of an instrument for inferring attentiveness on the basis of automatically collected emotional features and discusses the potential of the proposed method as well as its functionality and performance.

Chapter 5 gives a summary of what has been achieved through the study, namely, the identification of the needs pertaining to human motion recognition in actual environments and of the issues towards making such technologies possible, the proposal of methods for human motion recognition that address these different needs, and the verification of the effectiveness of these proposed methods. In addition, the study provided a classification of motion based on human intention and showed methods for inferring intention in accordance with the different types of motion. In particular, showing that it is possible to infer potential human interests through unintentional emotional behavior is considered as a very important achievement of the study. The study, therefore, has made a significant contribution in developing human motion recognition technologies that are useable in actual life environments, such as for creating new man-machine interfaces in the home, for personal profiling, and for detection of human behavior in surveillance videos and other images.

目次

図　目次

表　目次

第1章　序論

1.1　背景

近年では，個人が所有するＰＣや携帯端末にカメラが搭載されるようになり，映像を用いたコミュニケーションが一般的となった．またＴＶにもカメラ搭載モデルが登場しており，コミュニケーション用ツールとしてはもちろん，ユーザの状況把握に基づく表示制御機能までもが実装されている．インターネット上には大量の映像が溢れ，様々な映像を誰もが視聴できる環境が整っている．ゲームの世界ではMicrosoft Kinectの登場により，リモコンなしのジェスチャによるインタラクションが身近なものとなった．さらに街中にはいたるところに監視カメラが設置され，屋内外の映像を常時記録している．

このように，現代では映像が様々な形で日常生活の中に浸透しており，簡単にアクセス可能な身近なメディアとなっている．最近では，これら映像を人物動作認識のためのセンサとして利用することも盛んに検討されている．カメラによる非接触型センシングは心理的影響を除けばユーザへの負担が少なく，実環境での利用に適している．

映像解析による人物動作認識への社会的ニーズとして，たとえば動作をトリガとしたジェスチャリモコンがある．ＴＶをはじめとする現代の家庭内機器の制御は，通常赤外線リモコンによって行われている．赤外線リモコンは安価であり，遠隔操作できる利便性から，機器操作デバイスとしての地位を長い間保ち続けている．しかし一方でボタンの多さに対する不満や，紛失のしやすさ，機器の数に応じて増える台数などがユーザの不満を募らせている．デバイス不要の身体動作（ジェスチャ）でリモコン機能を代用できれば，これらの不満は解消できる．さらにジェスチャはリモコンとしての機能を超え，自然なコミュニケーションツールとしての活用も期待できる．

意図的な指示動作であるジェスチャのみならず，ユーザの無意図的で自然な動作を認識することへのニーズもある．たとえば「新聞を読む」，「うたた寝をする」などの日常動作を理解できれば，ユーザの状況を理解し，その状況に適した情報や機能をさりげなく提供できる．またユーザのわずかな動作や癖を計測し，集中度などの内部（心理）状態を推定できれば，ユーザ固有の趣味・嗜好の理解が可能となる．

映像コンテンツへの趣味・嗜好を理解することで，たとえばユーザ個別の番組推薦や情報提供サービスを実現できる．また不特定多数の視聴者の集中度を計測することで，そのＴＶ番組に対する質を測ることも可能になる．ＴＶ番組への評価は，視聴時間に応

じた“視聴率”が長年用いられているが，集中度の把握による“視聴質”を計測できれば，新たな番組評価尺度としての活用が期待できる．

また一般映像から人物行動を認識することへのニーズもある．放送映像や家庭用ビデオに“動作”に関するメタデータを付与できれば，人物行動に基づく映像検索が可能となる．動作のメタデータとしては，たとえば野球中継での“バッティング”，政治ニュースでの“握手”などがある．これら行動を自動検出できれば，ダイジェスト映像生成やスキップ再生などに応用できる．ただし一般映像からの行動認識は，撮影アングル，被写体のサイズ，非固定カメラ映像，背景ノイズ，オクルージョンなどの影響を考慮する必要があり，安定した解析は一般に困難である．

現在，行動認識へのニーズが最も高い分野の一つがセキュリティ分野である．近年では屋内，屋外問わずいたるところに防犯カメラや情報収集用カメラが設置されており，その数は増え続けている．しかし映像の確認は主に人手で行われており，非効率的な作業を強いられている．また市場調査や顧客行動パターンの解析に用いられることもあるが，こちらも自動解析しているケースは稀である．現状では監視カメラ映像は主に事後の検証用途で用いられているが，特定行動を自動・リアルタイムに検出できれば，犯罪の未然防止や即時の状況判断に活用できる．そのため，監視カメラ映像から特定行動や異常行動を自動検出する技術の早急な確立が求められている．

このように，映像解析による人物動作認識技術技術への社会的ニーズは高く，多岐にわたる．しかし制約条件のない実環境での映像を対象とした場合，前述の通り背景ノイズ, オクルージョン, 照明変化, 低画質映像など数々の撮影条件を考慮する必要がある．これら課題に対して頑健に機能する動作特徴量や認識処理が求められている．また人物動作は画一的でなく，その動作速度や動作量には個人差がある．さらに人物動作は動作者の意図の強さによっても多様に変化する．これら課題を全て解決することは非常に困難であり，人物動作認識技術が実環境へ応用された例はこれまでに少ない．

そこで本研究では，実環境でも有効に機能する映像解析技術，および多様な人物動作を頑健に認識する動作解析技術の検討を行った．前述の人物動作の多用性の問題から，あらゆる動作に対応した動作解析手法の確立は難しい．そこでまず人物動作を特定の基準で分類し，その分類に応じて個々の動作解析手法を検討することとした．

1.2　人物動作の分類

人体は多数の関節で成り立ち，それぞれが複雑な形状を持ち，関節同士が互いに影響しあいながら複雑な動作を構成している．人物動作を正確に理解しようとした場合，各関節とそれらの3次元位置変化を正確に求める必要がある．しかし2次元の映像情報から，

人体内部の骨格や関節の位置を正確に検出することは非常に困難である．

さらに動作にも様々な種類が存在し，それぞれの特性を考慮しなければ質の高い動作解析を行うことはできない．行動は動作の組み合わせとして構成されるが，動作解析を目指すうえでは，動作と行動を適切に分類することが必要である．しかしその分類は，研究の視点，調査の枠組みによって様々に異なる [Hirata97].

例えば，海保 [Kaiho00]は瞬間行動の分類として，技能ベース，規則ベース，知識ベースの3つに分類した．技能ベース行動とは刺激に駆動されて起こる行動であり，規則ベース行動とは目標にふさわしい動作が順次行われる行動である．また知識ベース行動とは，知識を動員して適切な判断を下した上の行動を指す．

久保田[Kubota06]は臨床行動心理学の見地から，図1.1に示すような運動・動作・行動の階層構造の内なる過程に着目した．階層構造によれば，“動作”は“運動”の組み合わせによって構成され，“行動”は“動作”の組み合わせによって構成される．これら3つの過程は互いに少しずつオーバーラップしているが，外からの観察でその流れを把握することはできると述べている．[Kubota06]では特に，人物の“動き”が動作者の意識や意志によって変化する点に注目した．

また加賀谷ら[Kagaya05]は，意図の強さにより身体動作を意図的動作（強情動を含む）と無意図的動作（情動）に分類した．特に無意図的動作（情動）を分析することで潜在的な感情を認識できるとし，情動に関して詳細な分析を行った．また意図的動作をジェスチャと行動全般に分類し，ジェスチャの分析で意思伝達動作を，行動全般の分析で動作の意図（行動理由・求めていること，感情）を理解できるとしている．

文部科学省の報告[Ministry of Education,Culture,Sports,Science & Technology in Japan05]によれば，“情動”とは「怒り・喜び・悲しみ・憎しみなどのような一時的な感情の動きで，表情，身振りなどの行動の変化や心拍数増加，血圧上昇などの自律神経系や内分泌系の変化を伴うもの」とされている．

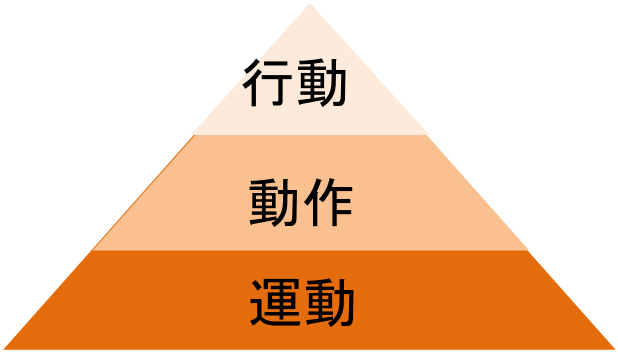


図1.1　動作の階層構造

松村[Matsumura06]は，情動を筋運動系，自律神経系，内分泌系の3つに分類した．筋運動系情動とは表情，態度，身振り，手振り，声などを表し，自身による調節（コントロール）が可能であるとしている．また自律神経系情動を瞳孔，発汗，心拍など，内分泌系情動をコルチゾール分泌であるとし，これら二つの情動をまとめて調節不能な情動性自律反応であるとした．3種の情動の中で，映像から確認可能なものは筋運動径情動のみである．筋運動系情動の中でも“表情”は動作者の意思による調節が容易であるが，“瞬目”や“視線変動”などの微小動作は比較的調節が難しく，動作者の内部状態を比較的真に表出すると考えられる．

上記知見をまとめ，意図の強さによって人物動作を分類した．その樹形図を図1.2に示す．まず，人間動作は意図の強い意図的動作と意図の弱い無意図的動作（情動）に大別される．意図的動作は，ある目的を果たすために腕や足を大きく動かす動作のことである．たとえば指を差す，走るなどがそれに相当する．これに対し，無意図的動作は動作者が意識せずに行う動作や身体変化のことである．たとえば首の傾げや瞬目がそれに相当する．意図的動作は無意図的動作に比べて視認性が高く，システム化は比較的容易である．通常，動作と言えば意図的動作を思い浮かべるが，実は意図的動作よりも無意図的動作の方が圧倒的に多くの割合で行われていると言われている．

意図的動作はさらに指示動作（ジェスチャ）と自然行動に大別される．そもそもジェスチャは相手に指示を出したり，情報伝達することを目的としており，その動きの視認性は非常に高い．そのため，ジェスチャ認識のシステム化は他の動作に比べて容易であ

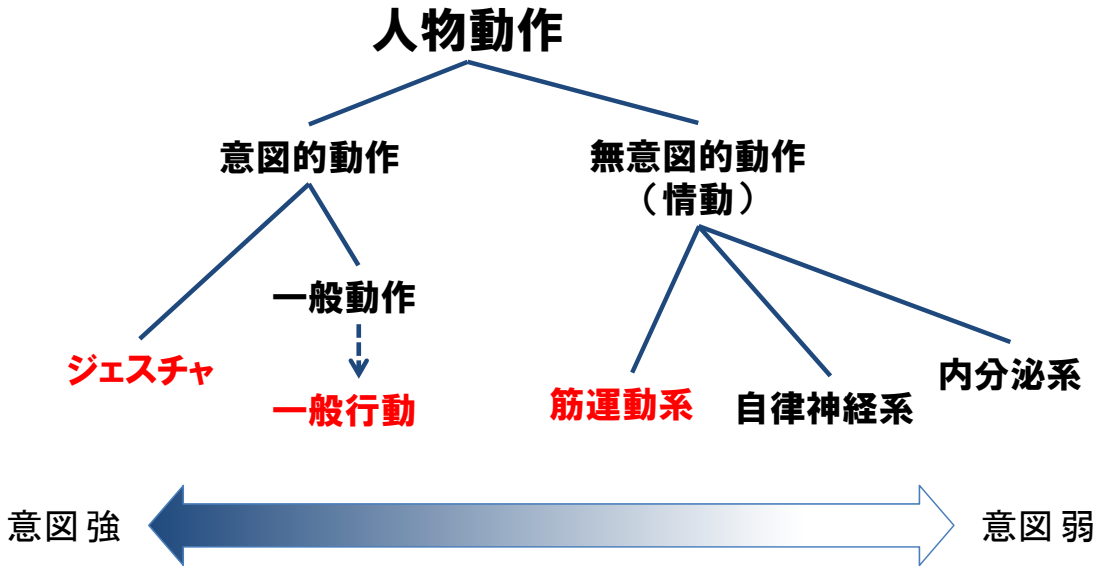


図1.2　意図の強さによる動作分類

る. 一方, 自然行動は他人への意思伝達を目的とせず, 自己完結するものが大半である. そのためジェスチャより視認性が低く，また動きの内容も多様であることから，システム化は一般に困難となる.

実際の研究進度においてもジェスチャの自動認識技術は比較的早い段階で研究が開始され，すでに実用化されているものも多い．しかし監視カメラから異常行動や特定行動を検出する技術はジェスチャ認識に比べて難易度が高く, 汎用的な実用レベルには到っていない. 画像の低レベル特徴解析による異常行動検出技術が実用化されている例もあるが，一般行動を認識しているとは言い難い．依然，多くの研究機関で自然な一般行動を認識するための特徴量や識別手法を検討しているのが現状である.

情動は一般行動よりさらに視認性が低く，システム化は困難を極める．前述の通り，情動は筋運動系，自律神経系，内分泌系に分類される[Matsumura06]．筋運動系の情動には顔の表情や身振り，手振り，声などの行動が含まれる．自律神経系の情動には瞳孔の収縮，精神性発汗，心拍数の増加などの生体信号が含まれる．内分泌系の情動に伴う変化には，副腎皮質から分泌される糖質コルチコイド（コルチゾール）分泌がある．このように情動でも筋運動系から内分泌系に至るまで視認性が減少する傾向にあり, 映像で確認できる情動は筋運動系までと考えられる.

情動に関する研究はその視認性の低さから, 脳科学や認知心理など科学的研究分野で接触型の専門機器を用いて解析されることが主であった．一方，情報学などの工学的研究分野で映像による情動解析が行われた例は稀であった. しかし筋運動系情動は可視情動であり，その動きは微小ながら映像による計測が可能である．昨今の映像機器の高性能化とともに，映像解析で情動を計測し，人間の内部状態を推定しようという試みも増えている.

本研究では上記課題を鑑み，主に次の2つの課題について取り組む.

1. 実環境での人物動作認識
2. 可視動作からの人物意思（内部状態）理解

課題1は，実環境に対応し得る人物動作認識技術を目指すものである．人物動作認識技術に対する社会的ニーズが高まっているが，実環境映像の撮影条件は多様であり，オクルージョンや背景ノイズ, 照明変化などの様々な影響を考慮した解析手法が必要となる．また家庭環境への応用を想定した場合，ユーザの負担を極力排除した形で人物動作をセンシングする必要がある．これら実環境下での課題を洗い出すとともに，制約条件のない実環境でも頑健に認識可能な人物動作認識技術を検討する.

動作認識とは，言い換えれば“動き”に表出する人物意思の理解とも言える．課題2

は，意図による人物動作分類を受け，分類した各動作の認識によって動作者の意思理解を目指すものである．図1.2では意図の強い順にジェスチャ，一般行動，筋運動系情動を示したが，意図が弱くなるに従い微小動作となり，視覚情報量が低下する傾向がある．本研究では，視覚情報のみから人物意思をどこまで自動認識できるかを検証する．さらには，映像コンテンツなどの外的要因の影響も考慮に入れ，動作者の内部状態を推定することを検討する．

1.3　一般行動認識技術の現状

映像解析による人物動作認識の研究は，既に長い歴史を持つ[Poppe10]．本項では，一般行動認識技術の現状についてまとめる．現状の課題や解決手段を例示することで，本研究を進める上での指針とする．

○　モデルベースとビジョンベース

人物行動認識技術は，人物骨格モデルの事前知識に基づいて人物姿勢を推定するモデルベースの手法と，画像の“見え”情報から直接人物姿勢を推定するビジョンベースの手法に大別される．モデルベースの手法は，身体部位の3次元位置情報を算出した上で姿勢を推定する場合が多い．単視点映像からでは身体部位の3次元位置を算出できないため，近年ではビジョンベースの手法で人物行動を認識するのが主流となっている．

映像から何らかの特徴量を抽出し，行動ラベルを付与するのが一般的なビジョンベースの動作認識アプローチである．クラス分類アルゴリズムは，通常トレーニングデータセットを用いて学習される．

○　人物行動認識を行う上での一般的課題

・　動作のクラス分類

多くの人物行動には膨大なバリエーションが存在する．例えば，同じ「歩く」行動にしても，速度や歩幅の長さによってその動作は変化する．また個人差も存在するため，人物行動のクラス数が増えるに従い，クラス間のオーバーラップが増加する傾向がある．人物行動認識手法の構築にあたっては，クラス内でのバリエーションを許容し，かつ他のクラスの行動を分離する手法が望ましい．

・　撮影環境

行動が行われている環境を考慮することも重要である．人ごみで混雑した環境下ではオクルージョンが生じる可能性があり，人物領域の抽出は困難となる．

また同じ行動であっても，撮影方向によっては見え方に大きな違いが生じる．複数のカメラを使用し，複数台分の映像特徴を1つの特徴記述子で記述できれば，特定視野でのオクルージョン問題は解決する．また背景領域の変化は，人物領域抽出や動作特徴抽出を妨げる要因となる．ムービングカメラで撮影した映像では，これらの影響は特に大きくなる．ビジョンベースの人物行動認識においては，これらの問題を確実に解決する必要がある．

- **時間変化**

一般に，動作は時間で分割されることが多い．人物動作を認識する前に，映像区間をセグメンテーションすることも人物動作認識の大きな課題である．

また動作者や状況によっては動作量が大きく変化する．動き特徴を用いて行動認識をする場合，動作量に不変な特徴量や認識アルゴリズムを用いる必要がある．

○　人物行動データセット

多くの研究は共通のデータセットを用いて評価されている．データセットの例としてKTH　human motion dataset, Weizmann human action dataset, INRIA XMAS multi-view dataset, UCF sports action dataset, Hollywood human action datasetなどがある．これらのデータセットは映画やWeb上の映像で構成されており，あらかじめ正解ラベルが付与されている．図1.3に人物動作データセットの例を示す．

図1.3　人物動作データセットの例（左列からKTH, Weizmann, Hollywood）

○　画像特徴

人物動作認識のための画像特徴は，理想的には少数の人物，背景，視点，動作から生成され，行動の分類や識別に有効な情報を保持していることが望ましい．

時間情報は動作を識別する上で非常に重要であり，動作認識を目的とした多くの画像特徴は時間軸を考慮している．一方，各フレームで個別に特徴を抽出する方法もある．この場合は，分類の段階で時間方向の変動を考慮する必要がある．

画像特徴は広域特徴と局所特徴に大別される．前者は画像や領域全体の“見え”に応じて記述される．広域特徴はトップダウン型の処理で抽出される．まず背景差分やトラッキングにより人物領域を抽出する．その後，注目領域全体を符号化し，特徴量として記述する．この記述子は多数の情報から生成されるため，強力な特徴量となり得る．しかし特徴量としての有効性は，人物領域抽出や背景差分，トラッキングなどの精度に依存するところが大きい．また，この特徴は視点やノイズ，オクルージョンによって過敏に変化する．これらの問題をうまく抑制できれば，広域特徴は有効に機能する．

局所特徴は，観測したオブジェクトを独立したパッチの集合体として記述する．局所特徴はボトムアップ型の処理で抽出される．まず時空間特徴点が検出され，局所パッチがそれらの特徴点の周辺で生成される．それらパッチは一つの記述子として統合され，特徴量として扱われる．局所特徴はノイズや部分的オクルージョンに比較的頑健であり，その有効性は背景差分やトラッキング精度に依存しない．しかし十分な量の特徴点を抽出する必要があるため，カメラの動作補正のような前処理が必要になる場合がある．

・　広域特徴

広域特徴は注目領域である人物領域全体を符号化する．人物領域は通常背景差分法やトラッキングにより抽出される．通常，広域特徴は2値化したシルエットやエッジ，オプティカルフローなどから生成されるが，ノイズや部分的オクルージョンや視点変化の影響を受けやすい．それら影響を軽減するため，グリッドベースアプローチなどが提案されている．また水平，垂直に加えて時間軸を考慮した時空間ボリュームなども利用されている．固定撮影カメラ映像の場合，人物のシルエットは背景差分法で所得できる．通常，実映像から抽出したシルエットにはノイズが含まれ，視点の影響を強く受ける．それでもシルエットは多くの情報を有するため，シルエットやその輪郭情報を用いた多くの手法が提案されている．

Bobickらは各フレームで作成したシルエットの変化領域を抽出し，Binary motion energy image (MEI)として活用した[Bobick01]．MEIには映像内で生じた動作の履歴が記録される．また直近の変化領域を高輝度で描画したMotion history imageも，シルエット画像を元にして生成される．

- グリッドベース広域特徴

注目領域を時間もしくは空間的に小さなグリッドに区切ることでノイズや部分的オクルージョン，視点変化などに対応させることが可能である．グリッド内の各セルは局所的な画像特徴を表している．このグリッドベース特徴は局所特徴に似ているが，注目領域の広域特徴に相当する．

Kellokumpuらは各グリッド内で前景領域の時空間LBPヒストグラムを計測した[Kellokumpu08]．またThurauらはHOGを利用し，前景領域のエッジに着目した動作認識を提案した[Thurau08].

- 時空間ボリューム

3次元時空間ボリューム(STV: spatio-temporal volume)は与えられたシーケンスでのフレーム内特徴を累積することで作成される．Blankらはまずシーケンス内のシルエット画像を累積し，STVを作成した[Blank05]．その後，ポワソンの方程式を用いて局所時空間顕著領域とオリエンテーション特徴を抽出した．これらの局所特徴を統合する重み付きモーメントを算出し，広域特徴とした．

多くの研究ではSTVをサンプリングし，その表面から局所特徴記述子を抽出している．このアプローチは局所特徴のアプローチと共通する部分が多数あるが，STVはあくまでも広域の特徴記述である．

・　局所特徴表現

局所特徴表現は，観測値を局所特徴記述子やパッチの集合として記述する手法である．そのため，高い精度の位置計測や背景差分処理は要求されない．また局所特徴は一定の視点変動，見えの変化，オクルージョンにも不変である．

各パッチは一定間隔，または時空間領域で動きに特徴がある点ごとに抽出される．局所特徴記述子は画像上の2次元小領域または映像内の3次元時空間領域から抽出される．広域特徴と同様に，観測値は部分領域でグルーピングされる．意味のあるパッチのみが保持されるため，パッチ内の時間と空間の関係性を考慮することで，人物動作をより効果的にモデル化することができる．

- Space-time interest point

時空間特徴点(STIP: Space-time interest point)は映像内で時間的・空間的に急激な変化が起きた位置を表す．LaptevらはHarrisのコーナー検出手法を3次元に拡張し，STIPとして活用した[Laptev03]．動きを考慮したこの特徴量は，人物行動認識に有効である．しかし，映像内で継続して行われる動作に関してはSTIPが生成されないため，長時間に及

ぶ動作の認識には適していない.

- 局所特徴記述子

局所特徴記述子は画像または映像パッチを要約して記述するものであり, 背景ノイズ, 見え, オクルージョンに不変であることが理想である. 回転やスケール変化に不変な特徴量がこれまでに提案されている. 時空間特徴のサイズは, 各方向成分における特徴点のサイズによって定められる.

局所特徴は単位の異なる異種特徴が混合していることが多く, また一般に高い次元を持っている. そのため, そのままの形で認識処理に用いることは難しい. そこで, 通常はコードブックが利用される. コードブックを用いることで, 抽出した各特徴をクラスタの中心か最も近いコードワードパッチへ量子化できる. この処理により, 局所特徴量は新たな記述子へと符号化され, 映像シーケンスはコードワードの頻度ヒストグラムで記述される.

- 局所特徴の関係性

局所特徴の関係性を考慮した特徴として, Scovannerらは共起ワードマトリックスを作成した[Scovanner07]. これはある動作は単独で起こることは少なく, 他の特定動作と共に起こりやすいという知見に基づいている. またこの手法はコードブックのサイズを軽量化できるという利点も兼ね備えている.

特徴点の追跡により, 局所特徴間の関係性を抽出することもできる. Sunらは各フレームでSTIP周辺のSIFT特徴を抽出し, 特徴点の軌跡を作成した[Sun09]. またMessingらはKLT法により特徴点軌跡を作成した[Messing09]. これらの手法は背景領域上のノイズの影響を受けやすく, 背景差分法などによる前景領域抽出が求められる. しかしこの軌跡特徴の利用により, 局所時空間特徴では取得できない長期にわたる動作特徴を取得できる.

1.4　本論文の構成と概要

次章以降の本論文の構成とその概要について以下に示す.

・　2章　次世代ＴＶ視聴環境における対話型ジェスチャ認識

次世代の大画面・高解像度ＴＶ視聴環境においては, リモコン機能を兼ね備えた新たなインタラクションツールとして, ジェスチャ認識が期待されている. 2章では意思伝達動作であるジェスチャに焦点をあて, 次世代ＴＶ環境に応用可能なジェスチャ認識手

法を検討する．はじめに，次世代ＴＶ環境で必要となるジェスチャについて検討する．次に，ジェスチャ認識における先行研究を紹介するとともに，本研究の目的を明確化する．そして長期の自然な対話型ジェスチャの認識に有効な，奥行き情報と局所軌跡特徴を利用した動作認識手法について述べる．最後に，提案手法を用いて試作したシステムの評価とその将来性について述べる．

- **3章　混雑映像を対象とした一般行動認識**

監視カメラ映像解析による人物行動認識技術に対する需要は高い．3章では人物の一般行動に焦点をあて，人混みで混雑した実環境での監視映像から，頑健に一般行動を認識する手法を検討する．まず関連研究により，一般行動認識技術の現状を紹介する．続いて広域特徴に基づく手法，局所特徴に基づく手法の2つの手法の比較により，実環境に適した行動認識手法を検討する．前者は人物追跡ベースによる手法であり，後者は局所軌跡特徴に基づく手法である．これらの検討を通じ，混雑映像下でも比較的頑健に人物の一般行動を検出するための手法を提案する．

- **4章　情動計測による注目度推定**

身体に表出する情動は動作者の内部状態を比較的正確に表出すると考えられる．4章では，映像コンテンツ視聴中のユーザの筋運動系情動を計測し，注目度に相当するユーザの内部状態を推定する．はじめに人間の内部状態推定に関する関連研究を紹介する．次に，複数の情動動作と注目度との関係性を示す仮説を立て，その妥当性を目視正解データにより検証する．そして各特徴を自動抽出する手法を提案し，その性能を評価する．最後に，自動計測した各情動から注目度を自動推定する手法を提案し，その性能評価を通して映像解析による人物の内部状態推定の可能性を示す．

5章　結論

最後に本論文で述べた研究成果についてまとめる．

第2章　次世代ＴＶ視聴環境における対話型ジェスチャ認識

2.1. はじめに

本章では，次世代のＴＶ視聴環境における新たなインタフェースを実現するジェスチャ認識手法を提案する．近年，ジェスチャ認識技術はPCやゲーム等の様々な家庭用機器に導入され，関連技術も多数報告されている．特にMicrosoft Kinectの登場により，ジェスチャ認識は学術的な研究対象としてだけではなく，商業的技術としても多くの関心を集めている．しかし，それら技術が認識し得るジェスチャの種類には限りがあり，指先を含む人物の自然な動作を理解できているとは言い難い．そこで対話型の自然な動作で機器とインタラクション可能なジェスチャ認識手法を検討した．

NHK放送技術研究所では次世代のテレビ規格として，ハイビジョンの16倍の解像度を持つスーパーハイビジョン(SHV)の実現を目指している．この大画面・高解像度ＴＶの登場により，従来とは異なるＴＶ視聴形態が生まれ，テレビとのインタラクション方法にも変化が訪れると予想されている．たとえば，拡大しても十分に高解像度であることから，好きな部分を拡大しながらのＴＶ視聴や，映像中に組み込まれた奥行き情報を利用してコンテンツ内オブジェクトとのインタラクションを行うことなどが考えられる．これら視聴中の画面操作や自然なインタラクションの実現にあたっては，リモコンなどのデバイスが不要であることが望ましい．そのため，テレビとの自然な対話を可能にする，ジェスチャによる新たなＴＶインタフェースの実現が期待されている．

そこで，映像解析技術による直感的所作で制御可能な，より便利でより自然なマンマシンインタフェースの開発を目指した．まず自然な対話型ジェスチャの認識を実現するため，奥行き情報を取得可能なセンサを検討した．本研究では，1台のセンサから近赤外線光を用いてオブジェクトの奥行きを測定するTime-of-flight（TOF）カメラの利用を検討した．

また従来のジェスチャ認識技術では，オプティカルフローなど時間的に局所的な特徴を用いている場合が多く，長期の自然な動作を認識することは困難であった．そこで本研究では，時間情報が豊富に含まれる軌跡特徴の利用を検討し，自然な対話型ジェスチャの認識を目指した．

提案システムの主な機能は，ジェスチャ認識とポインティング位置計測の2つである．ジェスチャ認識は，奥行きを含む映像情報を基に，様々な種類の人物ジェスチャを認識する機能である．まずTOFカメラの出力であるグレイスケール映像と奥行き映像から，局所特徴点を多数抽出する．次に，それらを追跡することで水平方向，垂直方向，時間方向，奥行き方向の4次元（4D）軌跡を作成する．ここで水平，および垂直方向の特徴点位置はグレイスケール画像から計測し，その特徴点の深度は奥行き画像から計測した．その後，これらの軌跡特徴をBag-of-features法（BoF法）に適用し，ユーザのジェスチャを頑健に認識した[Csurka04]．この4次元軌跡を利用したジェスチャ認識技術は本手法の主要技術であり，新規性を有する．従来技術との比較実験を通して本提案システムの有効性を確認した．

ポインティング位置計測は，カメラからユーザの顔および指先領域を検出し，その相対位置関係に基づいてポインティング位置を推定する機能である．奥行き情報の利用により，指先領域を頑健に抽出した．実際の操作を模したシミュレーション実験により，ポインティング位置計測機能の有効性を確認した．

本章で提案する新たなユーザインタフェースは，リモコンなどの接触型デバイスは不要である．デバイス不要でありながら，現状のモーションセンサコントローラのほぼ全ての機能を有することが特徴である．

本章では，まずユーザインタフェースを取り巻く技術の現状とともに，提案システムの特徴を紹介する．次に，ジェスチャ認識の従来手法について解説し，その課題と本提案手法の目的を示す．そして，奥行き情報と画像認識処理を組み合わせた新たな手法を提案し，その原理と仕様を詳述する．さらに，さまざまな実験を通して提案手法の有効性を確認する．最後に，本章で提案したジェスチャによる非接触型インタフェースの効果についてまとめる．

2.2.　次世代ＴＶ視聴環境でのユーザインタフェースの課題

人と機械の接点であるマンマシンインタフェースは，人間にとってより自然な方向へと常に進化してきた．PCにおけるテキストユーザインタフェース（TUI）からグラフィカルユーザインタフェース（GUI）への進化はその代表例であるが，最近ではキーボードやマウスすら不要のジェスチャによるユーザインタフェースの実現へと進化を遂げつつある．

一方，ＴＶのインタフェースは，ツマミの回転操作からボタン操作，そして赤外線による遠隔操作へと進化してきた．さらに現在ではスマートフォンなどのタブレット端末を用いたインタフェースも登場している．しかしこれらリモコンに対しては，「リモコ

ンの紛失」，「リモコンが増えて置き場所に困る」などの不満も多く聞かれる[Matsubara10].

そこで近年ではリモコン操作を身体ジェスチャに置き換え，体の動きでＴＶを操作する技術も開発されている．これらのシステムはＴＶに搭載したカメラの映像を解析し，ジェスチャを認識しているものが多い．しかし1台のカメラ映像からは被写体までの距離を算出できないため，自然な対話型ジェスチャの認識を実現できているものは少ない．

そのためステレオカメラなど，複数のカメラを用いてユーザへの距離情報を算出し，対話型ジェスチャ認識を実現している例もある[Rehg94, Nefian01]．しかし場所の制約などから，家庭環境で複数台のカメラを設置することは一般に難しい．実家庭においては，1台のＴＶ内蔵型カメラによるセンシングが望ましい．またテーブル型インタフェース等も提案されているが，対話型コミュニケーションのためにはＴＶと対面しながら操作できることが望ましい[Seifried09].

NHK放送技術研究所では，次世代のテレビ規格としてスーパーハイビジョン(SHV)の開発を進めている[Sugawara08]．これは現行のハイビジョン(HDTV)の16倍の解像度（7680×4320画素）を持つ高精細ＴＶであり，近い将来の放送開始を目指している．このスーパーハイビジョンが一般に普及すれば，ＴＶモニタのさらなる大画面化が進み，新たなＴＶ視聴スタイルの出現が予想される．図2.1に次世代ＴＶ視聴環境の概念図を示した．

ＴＶの大画面・高精細化により映像の立体感が増し，映像コンテンツへの没入感が高まる．そのため，ＴＶとのインタフェースにはデバイス不要のハンドジェスチャが期待されている．またSHVは高解像度ゆえ，画面の一部を切り出しても十分に高精細である．そのため，注視領域を拡大して視聴するスタイルが一般化すると考えられる．拡大視聴スタイルには，注視領域を微調整する平行移動操作も必要となる．ＴＶ視聴中にこのよ

図2.1　大画面・高解像度モニタでのＴＶ視聴

うなインタラクションを行うためには，デバイス不要のジェスチャによる操作が望ましい．

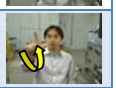
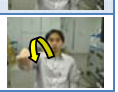
さらにユーザの没入感を妨げないよう，これら操作は指先の微妙な動きに合わせて実現されるべきである．単なる指示動作の認識ではなく，ユーザの意思を細やかに把握する機能が求められている．

また，オブジェクトの奥行き情報を含む新たな放送映像フォーマットが Moving Picture Experts Group（MPEG）と Society of Motion Picture and Television Engineers（SMPTE）で検討されている．映像に奥行き情報が畳重された場合，ユーザと映像内オブジェクトとのインタラクションが可能になる．没入感を損なわず，スムーズなインタラクションを実現するためには，ジェスチャによるポインティング機能が有効である．

あわせて，非操作動作による誤動作の抑制も必要となる．誤動作が起きれば，ユーザの没入感は損なわれる．操作・非操作のユーザ意図を理解するため，非操作動作を認識することも重要な課題である．

以上を鑑み，次世代ＴＶ視聴環境に必要なジェスチャを検討した．8種類の操作用ジェスチャと2種類の非操作用ジェスチャを表2.1に示す．

表2.1　次世代ＴＶ視聴環境でのインタラクションに有効なジェスチャ

ジェスチャ	機能	動作内容
Yes	決定	指の上下動 
No	キャンセル	指の左右動 
Zoom In	拡大/縮小	指の突き出し 
Zoom Out		指の引きよせ 
Left	位置調整	指を左に 
Right		指を右に 
Up		指を上に 
Down		指を下に 
Still	-	静止 
Move	-	ランダムな非操作動作 

まず“決定”と“キャンセル”動作を設定した．これらのジェスチャは，コンテンツ選択などの意思決定に必須である．またコンテンツの選択には，自由なカーソル移動も必要である．

現在このようなコンテンツ選択機能を実現するためには，リモートコントローラ，もしくはWii Remoteのようなモーションセンサコントローラが必要となる[Nintendo06]．これらデバイスからの計測値は，Bluetoothや赤外線を通してＴＶ側へ送られる．一方，提案手法でのコンテンツ選択は，ジェスチャ認識とポインティング位置計測機能により，接触型デバイスなしでの実現を目指した．

次に，“拡大・縮小”，“上下左右への移動”動作を設定した．前述の通り，次世代ＴＶ視聴環境では拡大・縮小しながらの映像視聴が一般的になると予想される．これらの操作は，この切り出し視聴に対応するものである．またこれらのジェスチャは没入感を損ねないよう，ユーザの微妙な腕・指先の動きと連動して行われるべきである．ユーザの意図を理解したジェスチャ認識を実現するためには，時間情報を豊富に含む特徴量の検討が必要となる．

最後に，非操作ジェスチャとして“静止”，“ランダム”を設定した．“静止”は動作が行われていないことを示し，“ランダム”は操作を意図していない大きな動作を示す．これらのジェスチャはシステムへの指示は与えないものの，視聴中の非操作ジェスチャの誤認識を抑制するためには必須である．

以上，次世代ＴＶ視聴環境でのユーザインタフェースの要件を示した．本研究での課題は以下の通りである．

1. 家庭用機器（ＴＶ）とのインタラクション
2. 長期的で自然な動作の認識

課題1の解決に向けては，奥行き方向の動作を認識可能なセンサを検討する．また実家庭環境で生じる背景ノイズの影響を軽減することも検討する．課題2の解決に向けては，ユーザ意思の強弱を理解するため，長期の時間情報を含む特徴量を検討する．

2.3. 関連研究

2.3.1 3次元位置情報の取得

ジェスチャ認識技術は，従来から多くの研究機関で研究されている[Hilton06]．認識対象ジェスチャはその用途によって異なり，指先・腕ジェスチャ，頭部および顔表情ジェスチャ，および全身ジェスチャの3種に大別される[Mitra07]．またジェスチャ識別の

ツールとしてはHidden Markov Models (HMM)，Particle Filter，Condensation, Finete-State Machine (FSM)などがある．これらの技術は主に単視点映像に対して適用されるものであるが，単視点映像から奥行き情報を取得することは不可能である．そのため，対話型ジェスチャの認識が必要な次世代ＴＶ視聴環境での入力インタフェースに用いることは難しい．

人物動作のセンシングデバイスとして，例えばelectromyographsやモーションセンサコントローラなどの各種センサが報告されている[Rajesh09, Bahar07]．これらのシステムは人物の位置，速度，及び加速度等の正確な値を感知して人物動作を高精度に認識することが可能である．しかし，それらデバイスの多くはその機能の複雑さから，家庭内での使用には適していない．特に高齢者の方の多くは，一般に普及しているリモコンですら使用に困難を感じている．さらに，複数人が同時に利用するためには人数分のデバイスが必要となる．そのため，映像解析による接触型デバイス不要なインタフェースへの需要は高い．

オブジェクトの奥行き情報は，人物行動やジェスチャを認識するためには非常に効果的である．人体部位の 3 次元位置情報を非接触型センサで取得する方法として，従来からステレオカメラが用いられている．これは被写体を視点の異なる 2 台のカメラで撮影し，2 枚の画像内の対応点の位置のずれ量を基に被写体までの距離を計測するものである．ステレオカメラによる奥行き計測は，そのシンプルな機構により特殊デバイスを用いることなく奥行き情報を取得できるため，多くの研究機関で長い間研究されてきた[Nefian01, Appenrodt10].

たとえば J. Rehg らは，ステレオカメラにより指先の各関節の 3 次元位置を計測し，3 次元空間でのマウス操作を実現するシステムを提案した [Rehg94]．指先のキネマティクスを考慮しており，非常に高速に動作することが特徴である．この手法以外にも，ステレオカメラによる姿勢推定・ジェスチャ認識は多数報告されている．

しかしステレオカメラは，奥行き情報を取得できる領域が両画像の対応点に依存するという問題がある．家庭での実環境映像に混入する背景ノイズの影響を避けるためには，画面全体の奥行き情報を取得することが望ましい．またステレオカメラは事前にカメラ較正が必要となる場合が多く，日常的な利用には適していない．その上，構造上2台以上のカメラが必須であり，装置が大掛かりになりやすい．上記の理由により，一般家庭でステレオカメラを利用することは困難である．

2010年，MicrosoftからKinectセンサが発売された[Microsoft10, Shotton11]．これは小さなセンサから無数の赤外線ドットパターンを照射し，赤外線カメラでそのパターンの投影状況を観測することで，被写体への奥行きを計測するデバイスである．この奥行き画像を基に，ユーザの各身体部位の3次元位置をリアルタイムで推定できる．このセンサ

はゲーム用途に開発され，身体ジェスチャを入力とした機器制御に成功した．

Kinectは1台の安価なセンサでユーザの動きを高速・高精度に計測できるため，ゲーム業界のみならずジェスチャ認識をはじめとするコンピュータビジョンの研究分野にも大きな影響を与えている[Xia11]. ただしこのセンサは人物の全身動作の理解が主目的であり，指先の細やかな動きまでは推定できない．そのため，次世代ＴＶリモコンとしての利用には課題が残る．

Time-Of-Flight (TOF)カメラもまた, 近年注目を集めている奥行き計測デバイスである[Ikemura10, Plagemann10, Ganapathi10]．これはセンサから赤外線を照射し，その反射光の到達時間から物体までの距離を計測するものである．計測結果を距離画像として，輝度画像と共にリアルタイム出力できる. また出力画像の各画素の距離情報を得られるため，カメラの前の人物領域を検出し，その全身運動を認識するだけでなく，指先のわずかな動きまでも計測できる．カメラ1台で運用でき，全画素での奥行き情報を計測できることから，ＴＶ視聴でのジェスチャ認識に有効なデバイスと考えられている．

松原らはこのTOFカメラを用いたＴＶインタフェースを開発した[Matsubara10]．このシステムは，ＴＶ視聴におけるほぼ全ての操作をユーザの手の振りや回転などの直観的ジェスチャによって実現した. またデザインもUIに配慮しており, 操作ストレスのない, 快適なＴＶ視聴を行うことができる．しかし，このシステムが認識する動作は短時間かつ離散的なジェスチャにとどまっており, 長期的で微妙な動作を認識するまでには至っていない．将来の大画面・高解像度ＴＶ環境での切り取り視聴を想定した場合，ユーザの意図を理解した長期的動作の認識が必要となる．

そこで，次に長期的動作の認識に有効な特徴量について検討した．映像解析による人物動作認識手法では，既に様々な特徴量が提案されている．次項では，人物動作認識に関する従来手法を調査し，長期の自然なジェスチャ認識に有効な特徴量を検討する．

2.3.2　人物動作認識のための特徴量

非接触型センサであるカメラは, 人物ジェスチャを認識するための入力デバイスとして有効である．しかし単視点映像からは奥行き情報を取得できないため，ジェスチャ認識用センサとしての地位は未だ確立できていない. ただし監視カメラ映像解析などの必要性から，単視点映像解析による人物行動認識の研究はこれまでに広く行われており，人間とコンピュータのインタラクション技術に関しても長く議論されている．

例えば，WRen らは「Pfinder」を提案した[Wren97]．これは人間の身体部位をリアルタイムで追跡するシステムである. また時空間情報を含む Motion Historical Image (MHI) が提案され，人物動作解析用途に長年利用されている[Valstar04, Ahad08]．さらに近年の監視カメラの急速な普及により，映像内の人物を追跡するだけでなく，特定行動を自動

検出することへの需要が高まっている. 特に通常行動から得られる画像特徴を多数学習しておき, それと異なる特徴を持つ異常行動を自動検出する技術が多数報告されている[Sillito08, Basharat09]. たとえば白木らは, 高次局所自己相関特徴を用いて映像から異常行動を検出する技術を提案した[Shiraki06].

ただしこれら従来技術は, 制約状況下で撮影されたビデオシーケンスを処理対象としていることがほとんどである. たとえば動作認識処理用の共通映像データセットとして有名な KTH, Weizman データセットでは, 平坦な背景上に大きく撮影された人物が比較的大きなジェスチャを行っている. KTH データセットの例を図 2.2 に示す.

これに対し, 実環境で撮影した映像シーケンスでは, 輝度変化やオクルージョンなどが頻繁に発生し, 安定した認識が非常に困難である. 制約条件のない実映像でも頑健な動作認識技術の提案が求められている.

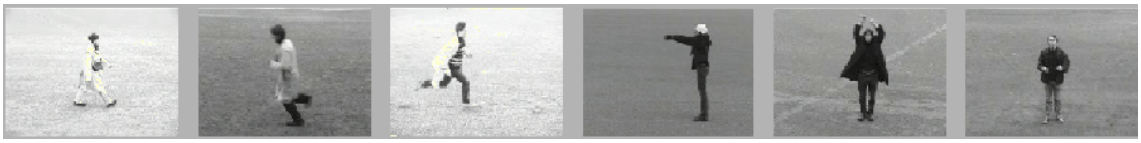


図 2.2　KTH データセットの例

Bag-of-features (BoF)法は, 画像内の局所特徴のヒストグラムで物体を認識する手法である[Csurka04]. 一般物体認識に適した技術であり, 局所特徴を利用しているためオクルージョンに頑健であることが特徴である. 言語処理分野では Bag-of-words 法として利用されてきたが, SIFT 特徴などの強力な局所画像特徴の出現により, 画像認識分野でも広く用いられるようになった. BoF 法は静止画での物体認識にとどまらず, 時空間特徴などをヒストグラム化することで, 人物行動認識にも応用されている[Schuldt04, Blank05, Fathi08, Sun09, Li08, Mikolajczyk08].

BoF 法を利用した人物動作認識のための特徴量として, 多くの画像特徴が提案されている. 例えば, Laptev は時空間特徴である Space-time interest points を提案し[Laptev03], Chen らは SIFT 特徴を時間方向に拡張した MoSIFT を提案した[Chen09]. これらは空間情報だけでなく時間情報を含んでおり, 人物動作解析に有効な特徴量である.

ただし上記特徴量の時間成分はオプティカルフローのような短期的特徴をベースとしている. そのため, 数フレーム内で収まる短期的行動はうまく記述できるものの, 数秒間にわたる長期的な行動を記述することは困難である. 将来のＴＶ視聴環境における長期的で自然なジェスチャの認識を想定した場合, これら特徴量は有効に機能しないことが予想される.

長期的な時間情報を有する画像特徴として, 軌跡特徴が提案されている. 軌跡特徴は

各フレームでの特徴点位置を記録した移動履歴であり，特徴点の位置情報が長期にわたり保持されている．そのため，長期的な人物動作を理解するために有効な特徴量として利用されている．例えば，Yoon らは掌で空中に書いた文字を認識するために軌跡特徴量を用いた[Yoon01]．この手法では，まず色情報と動き情報から画像内の掌領域を検出し，それを追跡することで掌の軌跡を作成する．その後，軌跡を時間方向にセグメンテーションし，各軌跡要素の位置，角度，速度特徴から書いた文字を認識する．認識に HMM を用いることで，掌の速度変化にも対応したことが特徴である．このように，軌跡特徴は長期の人物動作を理解する上で強力な特徴量である．ただし単一の軌跡を用いた動作解析は，その精度が対象領域の追跡精度に依存するため，背景ノイズの多い実環境ではうまく動作しないことが多い．

これに対し近年，特徴点軌跡を特徴に用いた BoF 法による動作認識手法が提案されている．Matikainen らは，画面内で抽出した多数の特徴点軌跡を BoF 法での features とし，人物行動を識別した[Matikainen09, Matikainen10]．この手法の特徴量は Trajectons と呼ばれ，オクルージョンにも頑健な性能を示している．長期的で自然なジェスチャ認識を実現するためには，この軌跡特徴を活用することが有効だと考えられる．

ただし，2 次元映像から得られる特徴は，水平方向と垂直方向の空間情報に限定されており，奥行き方向の動作を含むジェスチャの認識は難しい[Morency06]．そのため，対話型ジェスチャの認識に向けては，奥行きセンサの利用が有効であると考えられる．

本研究では以上の課題を鑑み，これまで別々に扱われてきたセンサ依存アプローチと特徴量依存アプローチを融合する新たな動作認識手法を提案する．具体的には，TOF カメラから得られる奥行き情報を用いて軌跡特徴を次元拡張し，4 次元軌跡特徴で人物ジェスチャを認識する．

奥行き情報を含んだ画像特徴は十分に検討されておらず，特に奥行き情報を加えた 4 次元特徴点軌跡によるジェスチャ認識手法はこれまでに提案されていない．本手法は，Matikainen らの手法[Matikainen09]を奥行き方向に次元拡張し，対話型ジェスチャ認識の実現と長期的で自然な動作認識の両面の解決を図るものである．提案手法は，全画像領域から多数の特徴点を検出・追跡し，抽出した 4 次元軌跡を特徴とした BoF 法で人物ジェスチャを認識する．したがってオブジェクトや身体部位の検出・認識を行うことなく，オクルージョンに頑健な動作認識処理を行うことが可能である．

一方で，ストレスのないマンマシンインタラクションの実現のため，ユーザが指し示している位置の計測も，解決すべき課題の 1 つである．デバイス不要のポインティング位置計測はジェスチャ認識と共に長い間研究されているものの[Park08, Nickel07]，奥行き情報の欠落による精度不足，煩雑なキャリブレーション作業の必要性など，様々な問題により実用的技術が未だ生まれていない．そこで，顔と指先領域の奥行き情報を活用

した，頑健なポインティング位置計測手法も提案する．提案手法は事前のキャリブレーションが一切不要であり，自然な動作でストレスなくポインティング操作を行うことができる．

2.4. システム概要

本研究が提案するジェスチャ認識システムは，入力部にあたる TOF カメラ，処理部にあたるプロセッサ，出力部にあたるモニタからなる．

TOF カメラは近赤外 LED 光の照射とその反射時間から，オブジェクトまでの距離を算出するデバイスである．反射時間の差に応じて物体までの奥行き値が画素単位で算出され，奥行き画像として出力される．TOF カメラは物体検出に優れた機能を有し，人物検出，身体部位認識およびマーカレスモーションキャプチャなどにも利用されている[Matsubara10, Ikemura10, Plagemann10]．近年では人物の動作認識デバイスとしても注目を集めているが，実際に実用化された例は未だ少ない．そこで TOF カメラデバイスのジェスチャ認識への応用を検討した．

本研究では TOF カメラに Panasonic 製 D-Imager を使用した[Panasonic09]．図 2.3 にその外観を示す．D-Imager は毎秒 12 フレームで作画し，1.2ー9.0 メートルの範囲内にあるオブジェクトの奥行きを計測することが可能である．奥行き画像のみならず，通常のグレイスケール画像も出力できる．両画像の光軸と視野角は同一であるため，オブジェクト内の特徴点は両画像で同じ位置に現れる．図 2.4 に TOF カメラからの出力であるグレイスケール画像と奥行き画像の例を示す．奥行き画像では，カメラに近い領域ほど高輝度で表示されている．

TOF カメラは現段階では比較的高価であり，一般的なデバイスとは言えない．しかし，その高い有効性により，将来的には安価で一般的なデバイスとなる可能性を十分に秘めている．たとえば，GPS やモーションセンサは数年前では珍しいデバイスであったが，現在では一般消費者にも広く普及しており，その精度も飛躍的に向上している．さらに，前述の通り MPEG と SMPTE で奥行き情報を含む新たな映像フォーマットが検討されている．放送映像に奥行き情報が組み込まれれば，奥行き計測への需要は高まり，距離センサは一般的なデバイスになると予想される．

開発に使用したプロセッサのスペックは次の通りである．CPU: 3.2GHz, メモリ: 4GBytes, HDD: 500 GBytes, OS: WindowsXP. また開発環境には Microsoft Visual C++を用いた．システムに含まれるモジュールはすべての独自に開発したが，特徴点検出や顔検出などの基本的なモジュールに関しては OpenCV ライブラリ[OpenCV00]を利用した．

本システムが提供するユーザインタフェースの処理フローを図 2.5 に示す．プロセッサへの入力は TOF カメラのグレイスケール画像と奥行き画像である．プロセッサは両画像を解析してユーザのジェスチャを認識し，またポインティング位置を計測する．姿勢やジェスチャに対するシステムの認識結果は，システムからの応答としてモニタ上に描画される．一方，ユーザはシステムからの応答に応じて指差しおよびジェスチャにより，システムへ指示を送る．本提案マンマシンコミュニケーションは，このフローを介して行われる．

図 2.3　TOF カメラ外観

グレイスケール画像　　奥行き画像

図2.4　TOFカメラからの出力

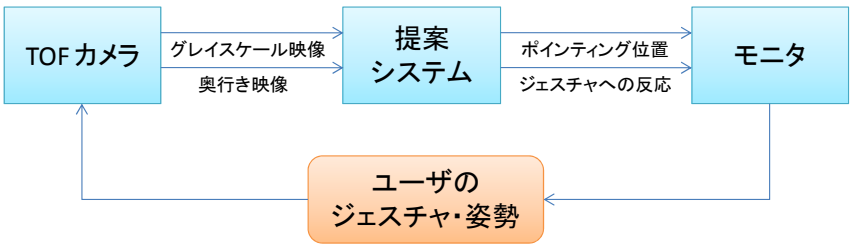


図2.5　提案手法のインタラクションフロー

2.5.　提案手法

提案手法は主にジェスチャ認識とポインティング位置推定の 2 つの機能を有する．ジェスチャ認識は“Yes”動作や“Left”動作など，ユーザの身振りによる指示を自動認識する機能である．ポインティング位置推定は，ユーザが指で差し示しているディスプレイ上の位置を自動計測する機能である．いずれも接触型デバイス不要の映像解析により実現される．

本章では，これら 2 つの機能について詳述する．

2.5.1.　ジェスチャ認識

本研究では，カメラ映像内の特徴点軌跡を特徴とした BoF 法によってジェスチャ認識を行う．BoF 法は局所特徴のヒストグラムで特徴を記述するため，オクルージョンに頑健であり，一般物体認識などに広く利用されている技術である[Csurka04]．本手法は人物領域周辺の特徴点の移動軌跡からジェスチャ認識に有効な特徴量記述子を生成し，BoF 法とサポートベクターマシン(SVM)でジェスチャを認識する．

ジェスチャ認識は学習フェーズと運用フェーズの 2 つのフェーズに分かれる．図 2.6 にジェスチャ認識の処理フローを示した．学習フェーズでは，まず TOF カメラのグレイスケール画像から特徴点を抽出・追跡し，奥行き画像の情報と統合して 4 次元軌跡を作成する．続いて抽出した無数の学習用 4 次元軌跡を *k*-means 法により *k* 個のクラスにクラスタリングする．そして各クラスタ中心を代表点とするコードブックを作成する．これら代表点はコードワードとも呼ばれる．

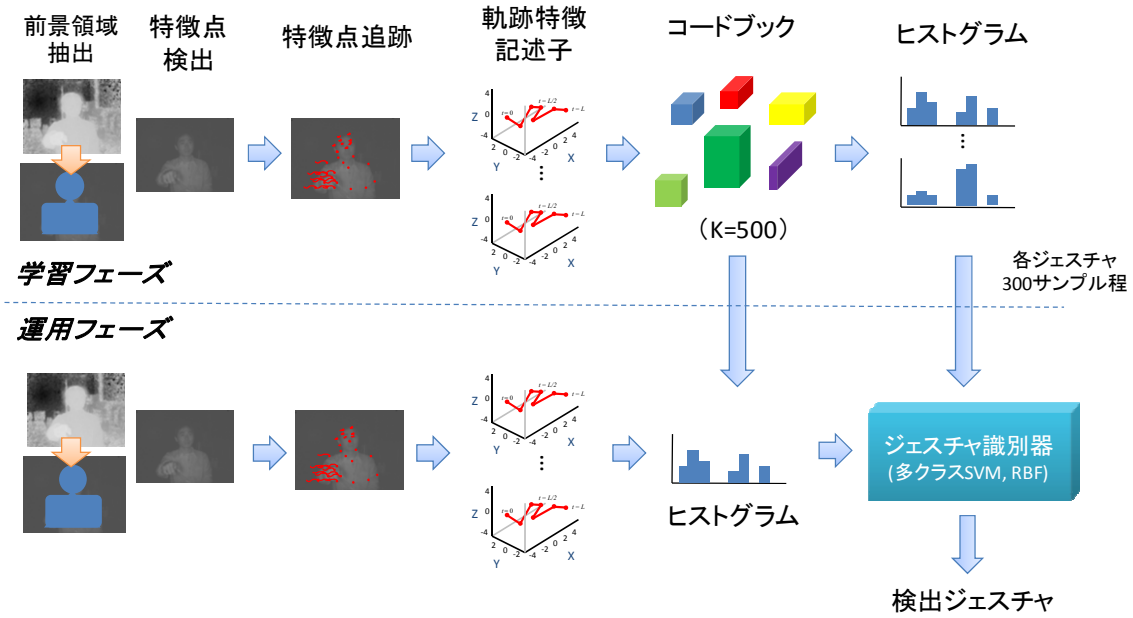


図2.6　ジェスチャ認識の処理フロー

次に，SVM [Schölkopf01, Chen05]によるジェスチャ識別器を作成する．学習用映像から抽出した軌跡特徴を k 個のコードワードの中で最も近いワードへ量子化し，k 個のビンを持つコードワードヒストグラムを作成する．このヒストグラムを最終的な特徴量として正解データと共に学習し，SVM 識別器を学習する．SVM のカーネルは RBF とした．

運用フェーズでは学習フェーズ同様，一定時間内で抽出した 4 次元軌跡特徴それぞれを最も近いコードワードへ量子化し，コードワードヒストグラムを生成する．このコードワードヒストグラムを基に，SVM 識別器でフレーム毎にジェスチャを識別する．規定時間内で過半数の得票を得たジェスチャをシステムの最終的な出力とした．

次項より各処理について述べる．

○ 特徴記述子

本手法では，ジェスチャ認識に 4 次元軌跡特徴を用いた．Matikainen らにより水平，垂直，時間の 3 次元特徴点軌跡を用いた人物動作認識手法が示されているが[Matikainen09]，奥行き方向も考慮した 4 次元軌跡特徴による人物動作認識は未だ提案されていない．本手法の特徴量は，Matikainen らの 3 次元軌跡特徴を奥行き方向に次元拡張したものである．4 次元軌跡特徴は 3 次元軌跡特徴より豊富な位置情報を有するため，高精度な動作認識が期待できる．

また奥行き情報で前景領域と背景領域を分離し，背景ノイズを排除した特徴点抽出を行ったことも本手法の特色の 1 つである．3 次元軌跡特徴と 4 次元軌跡特徴の概念図を図 2.7 に示す．図中の各点は隣接フレームでの特徴点の位置の差分を表し，L は軌跡の時間長を表している．

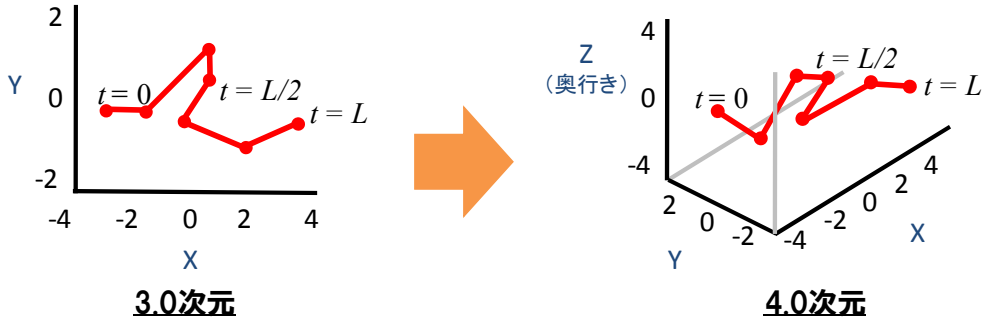


図2.7　Trajectons特徴量の次元拡張

4 次元軌跡特徴の生成手順を次に示す．TOF カメラからはグレイスケール画像と奥行き画像が出力される．まず奥行き画像を参照し，ユーザ領域（前景）と背景領域に領域を 2 分割する．前景領域だけを処理対象とすることで，背景領域上のノイズの影響を大幅に軽減した．続いてグレイスケール映像に Kanade Lucas Tomasi (KLT) トラッカ[Shi94]を適用し，前景領域内の特徴点を追跡する．KLT トラッカは映像内の特徴点を検出し，それらを追跡するアルゴリズムである．特徴点の検出には Harris オペレータを使用した．また特徴点追跡のためのオプティカルフロー検出には Lucas-Kanade の手法を用いた．KLT トラッカにより多数の特徴点が検出され，消失するまで十数フレームにわたり特徴点が追跡される．

続いて，得られた軌跡の各特徴点での 3 次元座標を算出する．まずグレイスケール画像から特徴点の水平，垂直位置座標(x, y)を観測する．次に奥行き画像上で同一座標に位置する画素値を参照し，奥行き座標(z)を観測する．奥行き座標 z には特定の補正量 w を乗じ，水平，垂直スケールとの正規化を施した．w は認識結果に大きな影響を与えなかったため，その値は実験的に定めた．

4 次元軌跡特徴作成の概要を図 2.8 に示す．本特徴量はセンサ依存アプローチと特徴量依存アプローチの二つの異なるアプローチからなる．従来ではそれぞれのアプローチが独立して研究されてきたが，それらを融合することにより，対話型ジェスチャ認識に適した特徴量を生成した．

4 次元特徴点の位置は $X_i^t = (x_i^t, y_i^t, z_i^t)$ と表わされる．ここで i は i 番目の特徴点を示し，t はフレーム番号を表す．各軌跡は，それぞれ L フレームの時間長を持つ軌跡片に分割される．

最後に，各特徴点の隣接フレーム間差分を算出し，4 次元特徴記述子を作成する．記述子の算出法を式(2-1)に示す．各軌跡片は最終的にこの $3 \times L$ 次元の特徴ベクトルとして表現される．

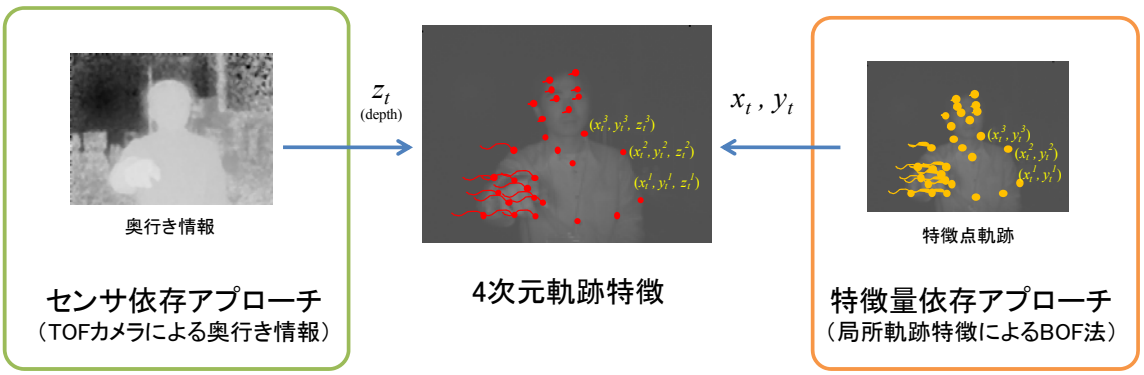


図2.8　4次元軌跡特徴の作成

$$\boldsymbol{T}_i^t = \{\boldsymbol{X}_i^t - \boldsymbol{X}_i^{t-1}, \boldsymbol{X}_i^{t-1} - \boldsymbol{X}_i^{t-2}, ..., \boldsymbol{X}_i^{t-L+1} - \boldsymbol{X}_i^{t-L}\} \quad (2\text{-}1)$$

○ 識別器の作成

学習フェーズでは，抽出した多数の軌跡特徴記述子にクラスタリング処理を施し，コードブックを作成する．まず，学習用軌跡特徴記述子をユークリッド距離を距離指標に用いた *k*-means 法により，*k* 個のクラスに分類する．その後，各クラスの中心をコードワードとしてコードブックへ登録する．コードブックに登録された軌跡特徴量記述子の例を図 2.9 に示す．これらの軌跡特徴量はジェスチャに含まれる典型的な軌跡であることを意味する．

新たに入力された軌跡特徴記述子は，作成したコードブックを参照して *k* 個のコードワードのうちの 1 つへ量子化される．得られた全ての軌跡特徴記述子にこの量子化を施し，*k* 個のビンを持つコードワードヒストグラムを作成する．軌跡数で除算し，度数の総和を 1.0 に正規化することで，フレーム毎に異なる軌跡数の影響を排除した．

このコードワードヒストグラムは，時間長 *L* の時間窓で抽出された軌跡特徴記述子から毎フレーム作成される．このコードワードヒストグラムを最終的な入力特徴とし，SVM 識別器でジェスチャを識別する [Schölkopf01, Chen05, Kurita08]．本手法では，RBF カーネルによる，1 vs 1 のマルチクラス SVM を作成した．また学習サンプルは，前述の 10 種類のジェスチャそれぞれで 300 サンプルとした．

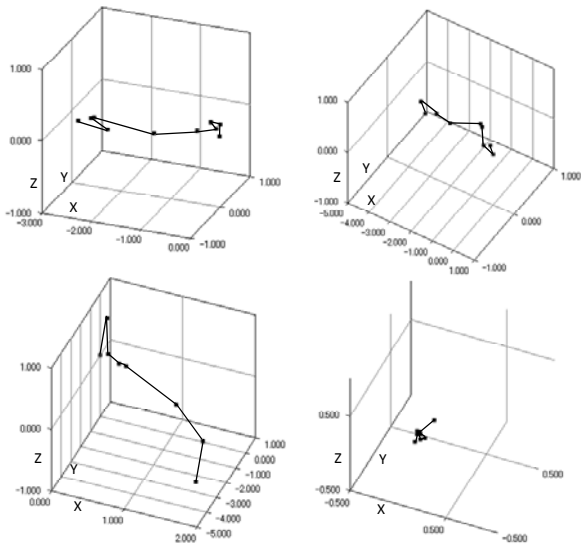


図2.9　特徴量記述子の例

○　ジェスチャ識別

前項で時間長 L の時間窓内にある軌跡特徴量を利用し，ジェスチャをフレーム毎に識別することを述べた．しかしフレーム毎の認識では，ノイズ軌跡の影響により識別結果が激しく変動する可能性もある．そこで，ジェスチャ識別の最終出力は時間窓 T フレーム内の投票によって定めることとした．

O_g の算出式を式（2-2）に示す．時間窓 T で過半数の得票を獲得したジェスチャ O_g を最終的な出力とした．ここで，p はジェスチャの ID を表し，V_p はジェスチャ p の得票数を表す．

$$O_g = \begin{cases} p: & p = \max_i V_i,\ and\ \ V_p > T/2 \\ null: & otherwise \end{cases} \tag{2-2}$$

ジェスチャ’Right’と’ZoomOut’の識別例を図 2.10，図 2.11 に示す．人物周辺の点と線は特徴点とその軌跡を表している．また左下の棒グラフは時間窓 T における各ジェスチャの得票率を表している．

時間窓 L 内の投票結果から最終出力を定めているため，認識結果は時間とともに徐々に他ジェスチャへと推移する．各ジェスチャへの投票割合の推移例を図 2.12 に示す．

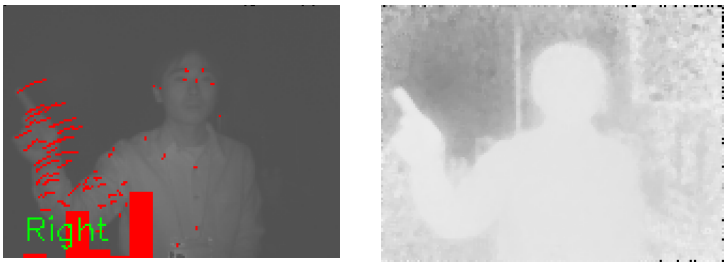


図2.10　“Right”ジェスチャの認識状況

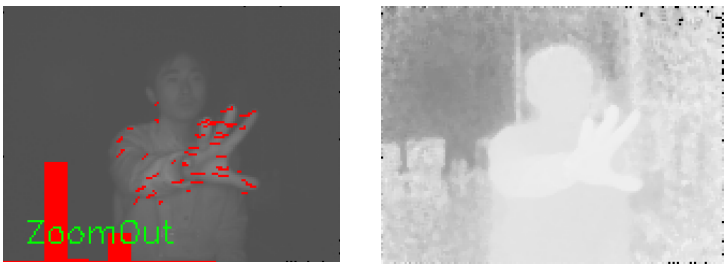


図2.11　“ZoomOut”ジェスチャの認識状況

図では，高い票を獲得したジェスチャを高輝度で示した．図の例では，’Still’ジェスチャからスタートし，開始後約 3 秒ほどで’Right’ジェスチャが検出され，約 6 秒ほどで’ZoomIn’ジェスチャが検出されている．この投票機能を実装することで，動作者の意図の強弱を考慮したジェスチャ認識が可能となった．

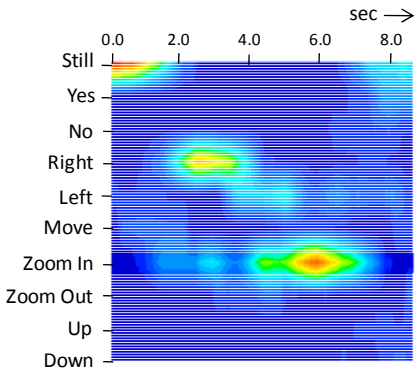


図2.12　得票率による認識ジェスチャの推移

2.5.2　ポインティング位置計測

提案手法は上記ジェスチャ認識機能の他に，ユーザのポインティング位置計測機能を備える．本機能は，ユーザの顔領域と指先領域の自動検出を行い，両者の位置関係からポインティング位置を推定する．本機能の処理フローを図 2.13 に示す．

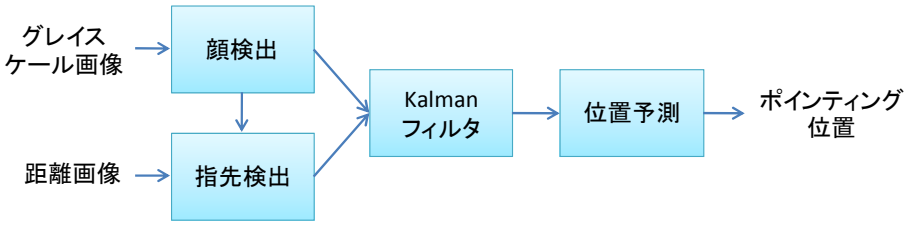


図2.13　ポインティング位置計測フロー

○　顔領域追跡

はじめに，グレイスケール画像中からユーザの顔を検出する．顔検出手法には Viola らの手法[Viola01]を利用した．この手法はその頑健性や高速性から顔検出に広く利用されており，オープンソースモジュールの OpenCV ライブラリにも含まれている[OpenCV00]．顔検出により，システムはグレイスケール上の顔領域中心(F_x, F_y)と顔領域のサイズF_sを得る．さらに，奥行き画像中で(F_x, F_y)の画素値を参照することにより，カ

メラから顔領域までの距離 F_z を得る．

顔領域の位置を毎フレーム確実に計測するため，Kalman フィルタ[Grimble94,Yu03]を利用した．Kalman フィルタはノイズを含む観測値の時系列データから，与えたダイナミクスに基づいてその真値を推定するアルゴリズムである．そのため，誤検出が生じても対象を滑らかに追跡できる．また観測が得られなくても真値を推定できるため，顔の検出に失敗した場合でも，その位置を推定することが可能である．本手法では，Kalman フィルタのダイナミクスに等速直線運動を用いた．

また顔検出処理を高速化するため，顔領域の探索範囲を限定した．Kalman フィルタはパラメータに誤差共分散行列 $\boldsymbol{P}$ を有し，その対角要素は，追跡が不安定な際に大きな値をとり，追跡が安定している際は小さな値をとる．そこで本手法では，$\boldsymbol{P}$ の値を参照しながら動的に探索範囲の大きさを定めた. 検出に失敗した際は $\boldsymbol{P}$ に応じて探索領域を拡大することで，頑健な顔領域追跡を実現した．

図 2.14 に顔領域の追跡例を示す．大きな円は検出されたユーザの顔領域とサイズを表し，その外側の矩形は次フレームでの顔領域探索範囲を示している．

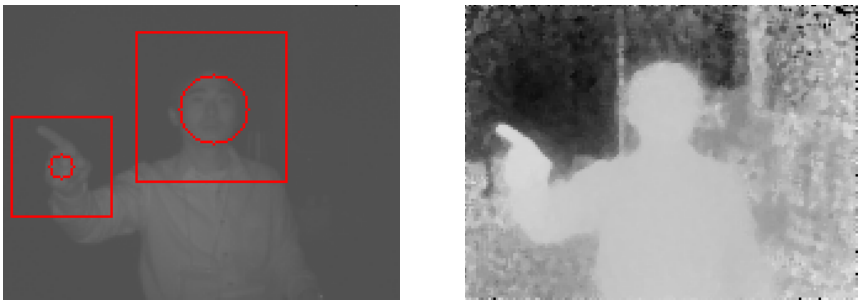


図2.14　顔領域検出・追跡と指先検出・追跡状況

○　指先領域追跡

顔領域を追跡した後，ユーザの指先領域を奥行き画像中から検出する．奥行き画像内で顔領域の深度 F_z よりも高い画素値を有する領域は，顔領域よりもカメラに近く，指先領域である可能性が高い．そこで，顔領域よりもカメラに近い領域を指先領域として検出した．そのような領域が複数検出された場合は，最もカメラに近い領域を指先領域とした．抽出した指先領域の中心位置(H_x, H_y)を指先位置とした．

顔領域同様，Kalman フィルタを用いて指先領域を追跡した．指先領域の形状はフレーム毎に変動するため，指先位置もフレーム毎に変動する．Kalman フィルタによる平滑化機能は，指先位置を安定して計測するためにも有効に機能した．図 2.14 に指先領域検出の例を示す．検出した指先領域を画面左の小円で示し，次フレームでの指先探索

領域を周辺の矩形で示した．

○　ポインティング位置推定

最後に，ユーザの顔領域と指先領域の相対位置関係より，ポインティング位置を推定する．相対位置 $\boldsymbol{R} = (R_x, R_y)$は式(2-3)により算出する．また顔領域と指先領域の相対位置の概念図を図 2.15 に示す．

$$R_x = \frac{H_x}{F_x}, \quad R_y = \frac{H_y}{F_y} \tag{2-3}$$

ユーザのポインティング位置 $\boldsymbol{P} = (P_x, P_y)$は式(2-4)にて算出する．$\boldsymbol{R}$ の最小値および最大値($\min R_x$, $\max R_x$, $\min R_y$, $\max R_y$)は事前の学習フェーズで計測しておく．

$$P_x = \frac{R_x - \min R_x}{\max R_x - \min R_x}, \quad P_y = \frac{R_y - \min R_y}{\max R_y - \min R_y} \tag{2-4}$$

P_x および P_y はそれぞれ 0.0～1.0 の範囲の値を持つ．最終出力としてのディスプレイ上でのカーソル位置は，P_x および P_y に使用しているディスプレイの解像度幅，高さを乗じて算出される．カーソル位置にも Kalman フィルタを適用しており，平滑化機能により滑らかな動きを実現した．

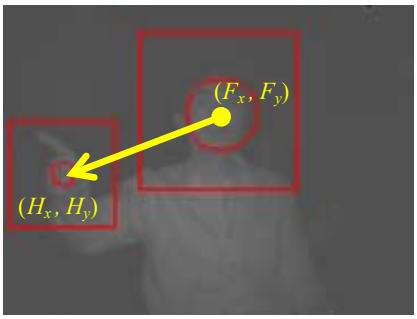


図2.15　顔領域と指先領域の位置関係

2.6.　実験

2.6.1.　実験条件

提案手法をシステムに実装した．本システムは図 2.16 のように TOF カメラ，プロセッサ, モニタの 3 つのデバイスからなる. 入力デバイスは TOF カメラ 1 台のみである.

また，モニタに表示される画面例を図 2.17 に示した．

ポインティング位置計測機能により，ユーザは自身の指先のみで映像コンテンツを選択できる．具体的には，ユーザがＴＶの前に指をかざすとキャラクターアイコンが現れ，その指がポインティング位置として機能する．ユーザが特定コンテンツのサムネイル上で一定時間姿勢を保持した場合，図 2.18 のようにコンテンツの詳細がポップアップウィンドウで表示される．また，初期化やキャリブレーション作業は一切必要なく，誰でも簡単に利用することができる点も本システムの利点である．

この提案システムの精度，汎用性，利便性を 4 種類の実験により検証した．パラメータとして，軌跡片の時間長 L=10[frame]，クラスタ数 k=500，出力平滑化のための時間窓 T=60[frame], SVM ソフトマージンパラメータ C=10.0 を用いた．

図2.16　提案システム外観

図2.17　コンテンツ選択画面の例

図2.18　コンテンツ選択中の画面例

2.6.2.　実験結果

○　4次元軌跡特徴の効果

はじめに，Matikainen らの手法[Matikainen09]との比較を通し，4 次元軌跡特徴の効果を検証した．実験では，1 人の被験者の操作用ジェスチャ 8 種類から軌跡特徴を抽出した. 同じジェスチャを 3 分間繰り返すことで, 学習用およびテスト用データを取得した. システムは 15 fps で動作するため，15[fps] × 180[sec]=2700[frame]のデータが得られた. 全データの 2/3 を学習用，残りの 1/3 をテスト用に用いた．本実験の 3 次元軌跡特徴は 4 次元軌跡特徴の奥行き座標を削除して作成したため, 両者の動きは完全に同一である.

図 2.19 に 8 種の操作用ジェスチャに対する 3 次元軌跡特徴と 4 次元軌跡特徴での認識結果を示す．縦棒は各ジェスチャの適合率を示し，棒上の線はその標準偏差の大きさを表している．この実験においては比較効果を明確に示すため，2.5.1 で述べた出力の平滑化を行わず，毎フレーム判定されるジェスチャをそのまま最終出力とした．そのため，対象ジェスチャ外のジェスチャが判定されることも多く，双方とも平滑化処理後の結果よりは低い精度となっている.

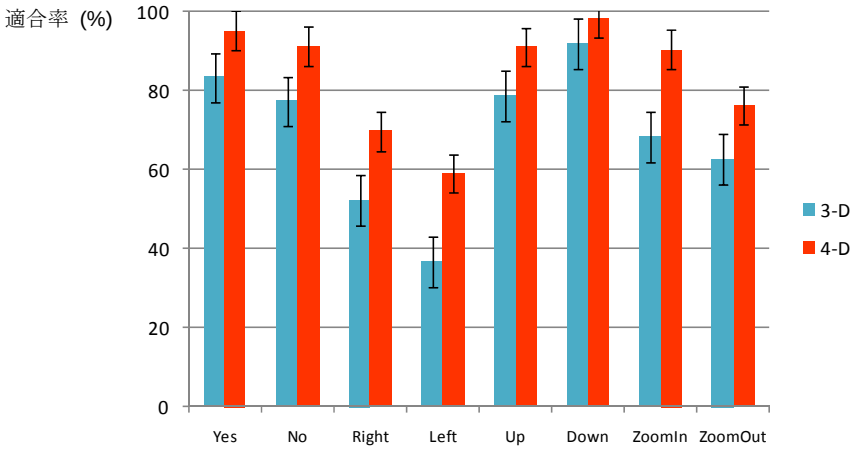


図2.19　3次元軌跡特徴と4次元軌跡特徴の比較

4 次元軌跡特徴によるジェスチャ適合率は 83.98%，3 次元軌跡特徴による適合率は 68.97%であった．奥行方向への次元拡張により，15.01%もの精度改善が得られた．*t*-検定による検証により，両者の結果には有意差が認められた．さらに，4 次元軌跡特徴による認識は，8 種類全てのジェスチャで 3 次元軌跡特徴による結果を上回った．これらの事実により，奥行き方向の利用による動作認識の有効性が示された．

‘ZoomIn,’ ‘ZoomOut,’ ‘Left,’ ‘Right’ジェスチャでは高い識別率の改善がみられた．特に‘ZoomIn’と‘Left’ジェスチャではその改善率が 20%を超えた．これらのジェスチャは奥行き方向の動きが顕著であるため，次元拡張が有効に機能したものと考えられる．

○　軌跡長と識別精度

続いて，Trajectons の軌跡長 L を 1, 3, 5, 10 フレームと段階的に伸ばし，軌跡長と識別精度の関係を評価した．図 2.20 にその結果を示す．実験条件は先の実験と同様である．

図より，軌跡長を延ばすほど識別精度が向上することが分かる．特に L=1 の場合は軌跡長が 1 フレームとなり，3 次元軌跡特徴量ではオプティカルフローに相当する．10 フレーム以上軌跡を伸ばした場合にはさらに精度が向上することが予測されるが，これは対象動作の平均時間に依存する．認識対象の動作に合わせ，軌跡長 L を適切に設定することが必要である．

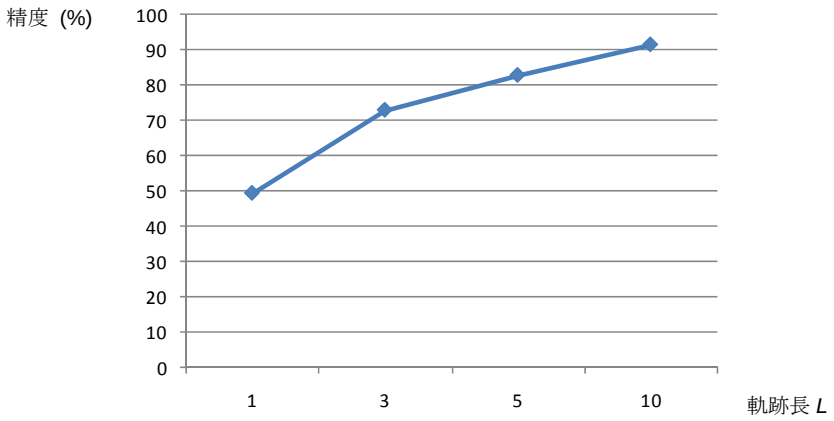


図 2.20　軌跡長と識別精度の関係

○　パラメータの検証

続いて，識別処理における各種パラメータの検証を行った．BoF 法でのコードワード数 k と識別精度，コードワード数と処理時間との関係を検証した．また，識別器の SVM のカーネルによる識別精度，ソフトマージンのパラメータ C と識別精度との関係を検証した．実験条件はいずれも先の実験と同様である．

・　**コードワード数**

図 2.21 にコードワード数 k と認識精度の関係を示す．コードワード数とは BoF 法における量子化での代表点数に相当し，コードブックのサイズとも呼ばれる．一般にこの値が大きいほど量子化誤差が減少するため識別性能は向上するが，その分識別時の距離計算量が増えるため処理速度は低下する．図より，コードワード数 700 で最も高い精度を確認した．ただしコードワード数 300 から 1,000 までの精度はほぼ均衡しており，500 程度コードワードがあれば識別性能上問題ないと考える．

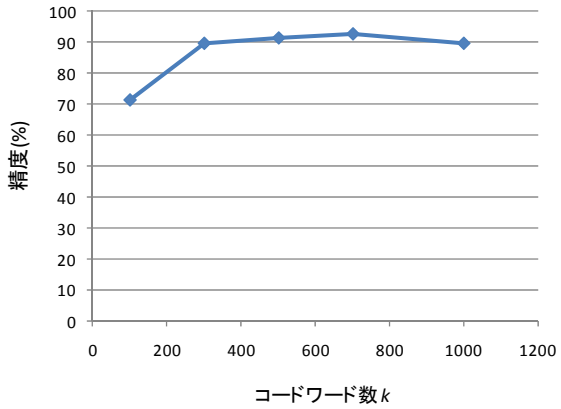


図 2.21　コードワード数 k と認識精度の関係

・　**コードワード数と処理時間**

図 2.22 にコードワード数と SVM による識別処理で生じた経過時間の関係を示す．図より，処理時間はコードワード数にほぼ比例して増加することが分かる．ただしコードワード数 700 を超えたあたりからやや勾配が急になっている．前実験でコードワード数 300 以上ではそれほど精度に差がないことが確認されているため，本手法のコードワード数は 500～700 程度が適当と考えられる．

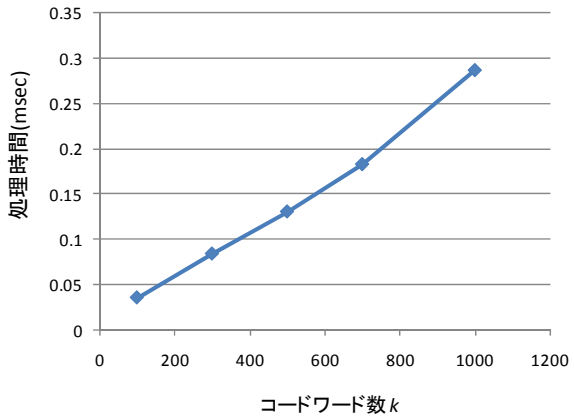


図 2.22　コードワード数と処理時間の関係

・　**SVM のカーネルと認識精度**

図 2.23 に SVM 識別器のカーネルと精度の関係を示す．ガウシアンカーネルとも呼ばれる RBF カーネルで最も高い精度を示した．RBF カーネルは一般に安定して動作することが確認されており，本システムにおいても有効に機能していることが確認できた．LINER カーネルは高次元特徴空間へのマッピングを行わず，元の状態空間で線形識別面を作成するカーネルである．最も高速なカーネルとして知られているが，その精度は一般に RBF カーネルよりも劣るとされており，本実験でも RBF カーネルの精度を下回った．

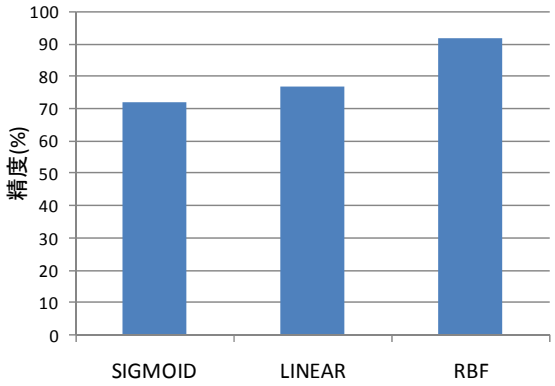


図 2.23　SVM でのカーネルと精度の関係

・　**SVM のソフトマージンパラメータ *C* と精度との関係**

図 2.24 に SVM のパラメータ *C* と精度との関係を示す．パラメータ *C* はソフトマージンを行う際のパラメータであり，マージンを破った判定を行った際の罰金に相当する．この値が少ないほど汎化性能は上がり，大きいほど学習データに沿った識別器が学習される．図より，*C* の値が大きくなるほど精度が向上し，C=10.0 程度で収束することが分かる．本手法では，この *C*=10.0 で識別器を学習した．

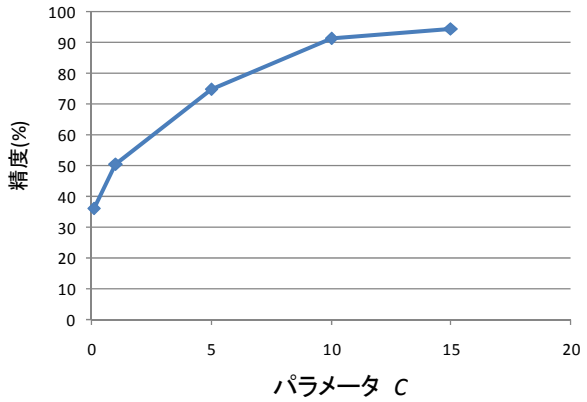


図 2.24　ソフトマージンパラメータと精度の関係

○　汎用性の検証

次に，複数人の被験者による汎用性の検証を行った．表 2.2，表 2.3 に 3 人の被験者による 30 回の各ジェスチャの認識結果を示した．識別器は 3 人の被験者以外のデータから作成したため，識別器に被験者の動作は学習されていない．この実験では，2.5.1 で述べた平滑化後の識別結果を用いて評価した．

表 2.2 は 4 次元軌跡特徴で識別したジェスチャのコンフュージョンマトリクスである．‘Right’ と‘ZoomIn’, ‘Left’ と‘ZoomOut’の間で多少の誤認識がみられた．この誤認識は，動作速度の個人差により生じたものと考えられる．しかしながら，総じて十分な識別率が得られており，4 次元軌跡特徴による識別は高い汎用性があると言える．

表2.2　4次元軌跡特徴量でのコンヒュージョンマトリクス

	Operational								Not Operational		
	Yes	No	Right	Left	Zoom In	Zoom Out	Up	Down	Still	Move	Not decided
Yes	1.00										
No		1.00									
Right			0.80		0.17						0.03
Left			0.03	0.87		0.10					
Zoom In					1.00						
Zoom Out					0.07	0.93					
Up							1.00				
Down								1.00			

表2.3　3次元軌跡特徴量でのコンヒュージョンマトリクス

	Operational								Not Operational		
	Yes	No	Right	Left	Zoom In	Zoom Out	Up	Down	Still	Move	Not decided
Yes	1.00										
No		1.00									
Right			0.63		0.30					0.03	0.03
Left				0.87		0.03					0.10
Zoom In					0.77	0.03				0.03	0.17
Zoom Out					0.07	0.87					0.07
Up							1.00				
Down								0.93			0.07

表 2.3 は 3 次元軌跡特徴を用いて識別した各ジェスチャのコンヒュージョンマトリクスである．表 2.3 における 3 次元軌跡特徴は，表 2.2 における 4 次元軌跡特徴から奥行き成分を削除して作成した．そのため，表 2.2 と表 2.3 で行われた動作は全く同一である．表 2.3 は，表 2.2 に比べて明らかに多くの誤認識がみられる．さらに，規定時間窓 *T* 内の処理で，どの操作用ジェスチャも過半数の得票を得られないケースも多くみられた．そのような場合は‘Not decided’に分類した．

最も多くの誤認識は‘Right’ と‘ZoomIn’ジェスチャ間で生じた．両ジェスチャでは 2 次元画像空間上で類似の動きベクトルが多数発生するため，3 次元軌跡特徴のみでの分離は困難であるためと考えられる．また‘Left’ と‘Right’で識別精度が異なるが，これは被験者が右手でジェスチャを行ったため，カメラからの視点では両動作が左右対称にならなかったためと考えられる．

‘Right’, ‘Left’, ‘ZoomIn’ , ‘ZoomOut’ジェスチャでは垂直方向に類似の動きベクトルが発生するため，頑健な識別が困難であった．そのため，それらのジェスチャの多くは過半数の得票を得られず，どの操作用ジェスチャにも分類されなかった．過半数の得票を得るまで時間を要するため，識別精度の低さは判定速度にも影響した．

表 2.2 の 4 次元軌跡特徴による平均適合率は 95.0%，表 2.3 の 3 次元軌跡特徴による平均適合率は 88.33%となり，4 次元軌跡特徴を用いることで 6.67%の精度改善がみられた．さらに，表 2.2 において 8 種の操作用ジェスチャ全ての適合率が 80.0%を超えた．この結果は奥行き情報の利用が識別精度の向上のみならず，識別可能なジェスチャの種類拡大にも有効であることを示している．さらに，4 次元軌跡特徴は判定時間の高速化にも貢献した．

○　シミュレーションによる操作性の検証

最後にシミュレーション実験により，本提案手法でストレスのない迅速な映像コンテンツ選択が可能かどうかを検証した．図 2.18 のような画面に 8 つのサムネイル画像を表示し，各コンテンツを選択するまでの時間を計測した．操作を困難にするため，各サムネイルを小さく，また隣接した位置に並べた．

コンテンツ選択操作は以下の 3 つの操作によってなされる．

- モニタの前に手をかざす
- カーソルを指先の操作で対象コンテンツのサムネイル上に移動する
- カーソルがサムネイル上にある状態で‘Yes’ジェスチャを行う

操作開始から終了までの時間を評価指標として計測した．被験者はコンテンツが選択されるまで’Yes’操作を繰り返すこととした．本実験では，6 人の被験者で合計 96 回の

試行を行った．結果を表 2.4 に示す．なお，被験者 A から D の 4 人の動作は学習データに含まれているが，E と F の 2 人の動作は学習されていない．

平均操作時間は 3.99 秒であり，その標準偏差は 0.24 秒であった．一方，市販のモーションセンサコントローラによる操作速度は 2.2 秒であった．市販のセンサと比較しても大きな遅延がなく，また 96 回の試行全てで誤操作は発生しなかったため，提案システムは十分な有効性があると考えられる．また学習の有無や被験者の違いにより，操作速度に有意な差は見られなかった．

最終的な操作時間に最も大きな影響を与えたのは，ジェスチャ認識処理であった．ジェスチャの判定はリアルタイムで毎フレームなされるものの，最終判定は一定時間窓 T 内での投票によってなされる．時間窓 T を短く設定することで最終判定までの時間を短縮できるが，その分誤認識も増加する．今後，認識率を向上することで，本手法での操作時間の短縮を検討したい．

本システムは，モーションコントローラによる操作より 2 秒程度長い操作時間を要する結果となったが，デバイス不要の利便性はその短所を補うものであると考える．

表2.4　コンテンツ選択作業の平均操作時間

Person	A	B	C	D	Average
Operation Time (sec)	4.03	3.83	3.88	4.22	3.99

2.7.　まとめ

将来の大画面・高解像度ＴＶ視聴環境を想定し，新たなＴＶインタフェースに必要なジェスチャ認識機能を検討した．本研究では操作用 8 種，非操作用 2 種の計 10 種類のジェスチャ操作を目指した．特に対話型の長期的な自然なジェスチャに焦点を当て，これらのジェスチャを頑健に認識可能な手法を検討した．

対話型ジェスチャの認識には奥行き情報の利用が有効である．本手法では入力に TOF カメラを用いて奥行き方向の情報を取得した．奥行き情報からユーザの顔や指先の 3 次元位置を計測することで，接触型デバイス不要のＴＶユーザインタフェースを実現した．

また長期的で自然なジェスチャ認識を実現するため，長期時間情報を有する特徴量の検討を行った．本研究では軌跡特徴量を用いることとし，TOF カメラで取得した奥行き情報から 4 次元軌跡特徴へと次元拡張した．4 次元軌跡特徴とは，水平方向，垂直方

向の空間情報，時間情報，そして奥行き方向の情報を含んだ特徴である．高次の軌跡特徴量による Bag-of-feautres 法により，多種の対話型ジェスチャを頑健に認識することが可能となった．

これまでの研究では，センサ依存アプローチと特徴量依存アプローチは独立に議論されてきた．TOF カメラと局所軌跡特徴の長所を用いることにより，両アプローチを融合した提案をしたことが，本研究の成果の一つである．

ジェスチャ認識に加え，ポインティング位置計測手法を提案した．ポインティング位置の計測にあたっては，顔検出技術とオブジェクト追跡技術を活用した．これら技術を用いてユーザの顔領域と指先領域を抽出・追跡し，両者の位置関係からポインティング位置を推定した．さらに Kalman フィルタによる平滑化処理により，カーソルの滑らかな移動を実現した．

各種実験を通し，本システムが次世代ＴＶ視聴環境で有効な 10 種のジェスチャを高精度で認識できることを確認した．8 種の操作用ジェスチャの適合率は全て 80%を超えた．また全ジェスチャにおいて，従来の 3 次元軌跡特徴による手法の精度を上回った．さらに，複数の被験者による実験を通してシステムの汎用性を確認した．

また TV コンテンツ選択を想定したシミュレーション実験により，提案手法がマンマシンインタフェースとして有効に機能することを確認した．既存のリモコンより短い操作時間は実現できなかったものの，デバイス不要の長所はその短所を補うものであると考える．

第3章　混雑映像を対象とした一般行動認識

3.1 はじめに

近年の計算機の高速化や映像解析技術の高度化に伴い，映像解析に基づく人物行動認識への期待が高まっている．行動心理学の定義[Kubota06]によれば，行動とは動作の有機的な組み合わせである．単純な「腕の上げ下げ」ではなく，「物を置く」，「指し示す」など，動作者の意図を含むより高次の動作と言える．そのため，人物行動の理解には動作の解析より複雑な解析手段が必要となる．

2 章で述べた意思伝達動作（ジェスチャ）よりも意図の小さい自然な動作，たとえば新聞を読む，うたた寝をする，などの認識が可能になれば，必要な情報を自動的にさりげなく提示するなど，より細やかな生活サポートを実現できる．また映像コンテンツに行動に関するメタデータを付与することで，野球のバッティングシーン，握手をするシーンなど，特定行動に関するシーケンスの検索にも利用できる．放送映像やインターネット映像コンテンツにこれら動作メタデータが付与されていれば，映像検索や自動要約等に有効に機能すると考えられる．

中でも人物行動認識が最も期待されているのはセキュリティ分野である．近年では街中のいたるところに防犯カメラが設置され，映像の情報量が増している．一方で映像を監視する側の人間の数には限りがあり，膨大な映像情報を有効に活用できていない．監視カメラの最大の利用目的はテロなどの犯罪を未然に防ぐことであるが，現状では犯罪後の検証に用いられていることが多い．

また，監視カメラは市場調査目的としても利用されている．商店の店内にカメラを設置し，混雑状況の確認や顧客行動パターンの検証などが行われている．しかしこれらも人間の目視確認による利用が主である．

これら監視カメラ映像を自動的に解析できれば，人件費の抑制だけでなく，防犯効率の向上やスピードアップにつながる．また店内の混雑状況の自動検出や顧客ニーズの把握にも有効である．

そのため近年では，単視点映像解析による人物行動認識の研究が盛んに行われている．しかし監視カメラ映像は一般に低画質・低解像度であることが多く，画像にノイズ成分を含むことが多い．野外では輝度変化が激しく，背景差分処理による安定した前景領域抽出は困難である．また樹木など，背景オブジェクトの影響を排除する必要もある．2 台のカメラで撮影していればステレオカメラの原理で奥行き情報を算出することも可

能であるが，特定の場所を 1 台のカメラで撮影している場合がほとんどであり，単視点の監視カメラから奥行き情報を取得することはできない．

さらに課題となるのが人物行動の多用性である．たとえば同じ「走る」行動を検出する場合でも，個人によってその速度や身体関節の使い方は異なる．さらに同一人物であっても，状況に応じてその所作は変化する．また各人の外見も様々であるため，人物領域の切り出しも困難である．

一般物体認識技術は SIFT や Haarlike 特徴と Bag-of-features 法を組み合わせることで精度の高い認識が実現しており，静止画像からのオブジェクト認識は実用段階に入っている．しかし映像からの一般行動認識では，空間情報のみならず時間情報を取り扱う必要がある．上記課題を克服するため，未だ各研究機関で有効な特徴量や識別手法を検討している段階である．

本章では，人混みで混雑した場所を撮影した単眼監視カメラ映像中から，人物行動を頑健に認識する手法を検討する．図 3.1 に本研究が対象とした混雑画像の例を示す．本章ではこのような映像で大量に発生する激しいオクルージョン，人物の多様な外見，動作の個人差などの課題に対処し得る頑健な人物行動認識技術を検討する．

図 3.1　混雑画像の例（TRECVID データセット）

3.2　関連研究

映像解析に基づく人物行動認識技術は, 多くの研究機関で長年にわたり研究されている．人物行動認識技術は，人物骨格モデルの事前知識に基づいて人物姿勢を推定するモデルベースの手法，画像の“見え”情報から直接人物姿勢を推定するビジョンベースの手法に大別される．モデルベースの手法は，身体部位の3次元位置情報を算出した上で姿勢を推定する場合が多い．しかし単視点映像からでは身体部位の3次元位置を算出できないため, 近年ではビジョンベースの手法で人物行動を認識する手法が主流となっている[Poppe10].

時間情報は動作を識別する上で非常に重要であり, 人物動作解析における多くの画像特徴は時間情報を含んでいる．一方，各フレームで個別に特徴を抽出する方法もある．この場合，分類フェーズにて時間方向の変動を考慮する必要がある．

画像特徴はグローバル特徴と局所特徴に大別される. 前者は画像や領域全体の“見え”に応じて記述される．グローバル特徴はトップダウン型で取得される．まず背景差分やトラッキングによって人物領域を抽出する．その後，注目領域全体を符号化し，特徴量として記述する．この記述子は多数の情報から生成されるため，強力である．しかし人物領域抽出や背景差分，トラッキングの精度に依存するところが大きい．また，この特徴は視点やノイズ，オクルージョンによって過敏に変化する．これらの要素をよく制御できれば，通常グローバル特徴は有効に機能する．

Gorelick らは，時空間ボリュームと呼ばれるグローバル特徴を用いて人物行動の認識を行った [Gorelick07]．時空間ボリュームとは，固定カメラ映像中のオブジェクトの領域変化を累積したものである．映像の較正が不要，周期的行動以外も認識可能，部分的オクルージョンにも対応可能，などの特徴を有する．この先行研究では，時空間ボリュームにより 10 種類の人物行動を検出している．

また Dalal らは，グリッドベースのグローバル特徴である Histograms of oriented gradients (HOG)特徴を提案し, 映像中から人物領域を抽出する手法を提案した [Dalal05]. この手法は部分的オクルージョンに頑健であり, その高い精度から人物領域検出のスタンダードとなっている[Beiping11]．また藤吉は，この HOG 特徴を用いて人物領域を検出し, 隣接フレームでの同一人物の位置を連結することで人物の移動軌跡を自動作成した [Fujiyoshi07].

ただし，これまで提案された動作認識手法の多くは，図 2.2 で示した KTH[Schuldt04] や Weizmann[Blank05]のように，比較的平坦な背景上で 1 人もしくは少人数が動作する映像データセットを解析対象としている[Fathi08, Sun09]．混雑した群衆映像を対象とした研究も行われているが, その目的は人物追跡や群衆単位の動作認識などに限定されており，混雑した実映像中の個人行動を解析した例は少ない[Lien08, Tsai06, Hu08].

人物行動認識技術の実用化を目指す上では，混雑した実映像で認識手法を評価することが必要である．近年，そのような課題に取り組む試みとして，TRECVID に代表される評価型ワークショップなどで，汎用性の高い動作認識手法の評価が行われている．

混雑映像の解析にはオクルージョン対策が必須であり，局所特徴が有効に機能すると予想される．この局所特徴はパッチとも呼ばれ，ボトムアップ型で取得される．まず時空間特徴点を検出し，続いて特徴点周辺で局所パッチを算出する．最後に，対象シーケンスを独立したパッチの集合体として記述する．

局所特徴はノイズや部分的オクルージョンに比較的頑健であり，背景差分やトラッキングの精度に依存しない．しかしながら十分な量の特徴点を抽出する必要があるため，カメラの動作補正のような前処理が必要になる場合がある．

時空間領域での局所特徴の例として，たとえば laptev は Space-time interest points (STIP) を提案した[Laptev03]．また Chen らは MoSIFT と呼ばれる特徴量記述子を提案した[Chen09]．これらの特徴量は空間情報のみならず時間情報も保持しているため，人物行動の認識に適している．また局所特徴は部分的オクルージョンに頑健であり，混雑した映像解析に用いる特徴量としては有効である．しかし，これらの特徴量の保持する時間長は数フレーム程度であり，「走る」などの長時間にわたる行動の検出は難しい．長時間行動の検出には，時間に関する情報が豊富な特徴量が必要となる．

特徴点軌跡は特徴点の出現から消失までの位置座標を記録しており，特徴点の“動き”に関する比較的長期の情報が保持されている．特に人物領域周辺で抽出される特徴点軌跡には，人物動作に関する多くの情報が含まれている．そのため，特徴点軌跡を用いた人物動作認識手法がこれまでにも提案されている[Matikainen09, Matikainen10, Li08, Perbet09, Bhuyan08].

Bag-of-Words 法は文書内の単語の頻度ヒストグラムから文書のトピックを推定する手法であり，言語処理分野で広く用いられている．SIFT などの強力な局所画像特徴の誕生により，画像処理分野にもこの手法が応用されるようになり，Bag-of-VisualWords 法，Bag-of-features 法（BoF 法）などと呼ばれている[Csurka04]．オクルージョンに対する頑健性から，とくに一般物体認識に有効な手法として知られている．

人物行動認識においても，この BoF 法を利用した研究が盛んに行われている．しかし特徴点軌跡はそれぞれに異なる時間長を持つため，固定次元の特徴量を想定した BoF 法へそのまま利用することは難しい．そのため軌跡を固定時間長で区切るなど，軌跡特徴を固定次元化する処理が必要となる．また通常の BoF 法は各特徴の重みを考慮しないため，背景に含まれるノイズ特徴の影響を受けやすいという課題もある．

近年，時間長の異なる特徴点軌跡から移動方向に関する bin 数固定のヒストグラムを作成する手法が提案されている[Mezaris10, Sun09]．軌跡の時間長や位置，大きさに不変

な特徴量であり，動作者の速度や位置の変化にも頑健である．軌跡ヒストグラム特徴と BoF 法の組み合わせにより，比較的頑健に人物動作を認識することが可能である．

ただし，BoF 法に起因するノイズ特徴の影響については十分に検討されておらず，上記手法は改善の余地があると考えられる．実画像解析においては一般に大量のノイズが特徴量へ混入する．これらのノイズから抽出した特徴量は前景領域上の特徴量と同じ重みで扱われるため，大きな問題となる．自然言語処理で用いられることの多い *tf-idf* 法は，特徴量の重みを算出する有用な手段である[Salton88]．この *tf-idf* 法で各軌跡の重要度を算出し，重みとして付与することで，より頑健な人物行動認識が期待できる．

また局所特徴の BoF 法による解析手法では，画面全体を 1 つの Bag として処理することが多く，画面内に複数の人物がいるときには人物単位の処理を行うことができない．この課題に対処するため，軌跡特徴をクラスタリングする手法が存在する．杉村らは，混雑映像中から多数の特徴点軌跡を抽出し，それらをクラスタリングすることで人物単位の領域追跡を実現した[Sugimura10]．この手法では，歩容特徴（人物固有の歩行周期）と局所的な見え（3 角パッチ内の時間変化）の一貫性を利用して人物領域を追跡した．背景領域に含まれるノイズ軌跡の影響も軽減でき，また人物非検出のためオクルージョンにも頑健である．

上記の先行研究を踏まえ，次項では混雑映像から人物行動を頑健に検出する新たな手法を提案する．

3.3　提案手法

3.3.1　提案手法の概要

本研究では，解析困難な TRECVID の映像データセットをもとに研究を進めた．TRECVID は米国 NIST が主催する映像検索に関する国際的評価型ワークショップである[NIST, Smeaton06]．毎年世界各国の研究機関が共通のデータセットを用いて共通のタスクに挑んでいる．実映像を解析対象とし，非常に困難なタスクが設定されることが TRECVID の特徴である．

TRECVID Surveillance Event Detection(SED)タスクは空港に設置した 5 台の監視カメラ映像から 7 種類の特定行動（走る，物を置く，指を差す，人に会う，人と別れる，携帯電話をかける，抱擁する）のうち，3 種類以上を自動検出するタスクである．図 3.1 のように，非常に混雑した空港内を撮影した映像であり，ロバストな解析が非常に困難である．我々はこの TRECVID への取り組みを通し，遮蔽やノイズに頑健な人物動作認識手法を検討した．

本研究では，大きく分けて次の 2 つの人物行動認識手法を検討した．

手法 1：　人物追跡ベース

手法 2：　特徴点軌跡ベース

手法 1 の人物追跡ベースでは，グリッドベースの広域特徴である HOG 特徴を用いて映像中の人物領域を検出し，その移動軌跡から人物行動を認識した．図 3.2 に人物領域の検出例を示す．人物行動認識はあくまで人物の動きを解析する技術であり，前処理で人物の検出が必要であるという思想に基づく手法である．

この手法のメリットとして，人物検出により背景ノイズの影響を受けずに人物動作を認識できる点があげられる．隣接フレームでの人物検出位置を連結することで，図 3.3 のような人物軌跡を作成できる．この軌跡を分析すれば，「走る」，「反対方向へ歩く」など人物の移動を伴う大きな行動の認識が可能になると予測される．

図 3.2　人物領域検出の例

図 3.3　人物追跡により作成した人物軌跡

一方でこの手法によるデメリットも存在する．図 3.4（左）に示す「物を置く」などの比較的小さな行動では人物の移動を伴わず，移動軌跡の解析では検出が困難であると考えられる．また，図 3.4（右）のように非常に混雑した映像においてはオクルージョンの影響で人物一人一人を正確に検出することは難しく，認識精度が低下することも予想される．

図 3.4　移動を伴わない動作の例（左）と極度に混雑した映像例（右）

手法 2 の特徴点軌跡ベースは，局所特徴である特徴点の追跡により多数の特徴点軌跡を作成し，その特徴を用いた BoF 法により人物行動を認識する手法である．図 3.5 に特徴点軌跡の例を示す．この方法では人物検出を行わないため，人物行動認識の精度が人物検出の精度に依存しない．そのためオクルージョンが多い混雑映像に適した手法と考えられる．しかしながらこの手法では人物単位の処理が難しく，背景ノイズの影響を受けて大量に False Alarm が発生することも予想される．

図 3.5　特徴点軌跡の例

この手法のメリットは，前述の通りオクルージョンに頑健である点にある．また図 3.6 に示すような「携帯電話をかける」などの移動を伴わない動作も頑健に検出できるため，より汎用的な人物行動認識手法であると考えられる．

一方のデメリットとしては画面全体を一つの Bag として処理するため，背景ノイズに弱い点があげられる．図 3.7（左）に背景ノイズの例を示した．また，画面奥の人物からは少量の軌跡しか得られないため，画面奥の人物行動の検出もれが生じるおそれもある．図 3.7（右）に画面奥で「走る」行動が発生している例を示す．

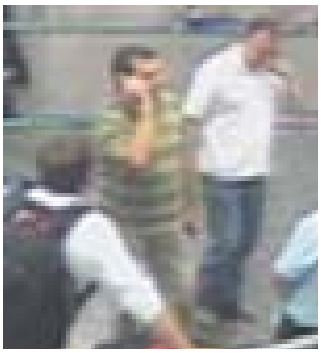


図 3.6　小行動（携帯電話をかける）の例

図 3.7　背景ノイズの例（左）と画面奥での「走る」行動例（右）

さらに手法 2 の発展版として，軌跡特徴のクラスタリングに基づく手法を検討した．手法 2 では人物単位の処理が行えず，背景ノイズの影響を受けることが予想された．そこで軌跡特徴集合に互いの距離を基準としたクラスタリング処理を施した．これにより，各クラスタで独立した行動認識が可能となった．軌跡のクラスタリングを施した例を図 3.8 に示す．図中の楕円が一つのクラスタを示している．

本研究では主に上記 2 つの手法の比較を通し，混雑した映像においても有効に機能する人物行動認識手法を検討した．

図 3.8　軌跡のクラスタリングの例

3.3.2　人物追跡ベースの手法（手法 1）

○　手法1の概要

本項では人物追跡ベースの手法の概要を示す．本手法への入力は，リピート再生や過去方向への再生が可能な映像データとする．入力映像から検出された特定行動の映像区間が解析結果として出力される．提案手法の概要を図 3.9 に示す．

本手法は学習と運用の 2 つの処理フェーズで構成される．学習フェーズでは学習用映像データセットを用いて人物検出器を学習し，また行動認識に必要な各種パラメータを設定する．学習処理は対象カメラ毎に行う必要があるが，本手法は単眼カメラ映像を対象としているため，カメラキャリブレーションなどの作業は一切不要である．運用フェーズでは，評価用映像データセットから軌跡特徴量を抽出し，特定行動を自動検出する．

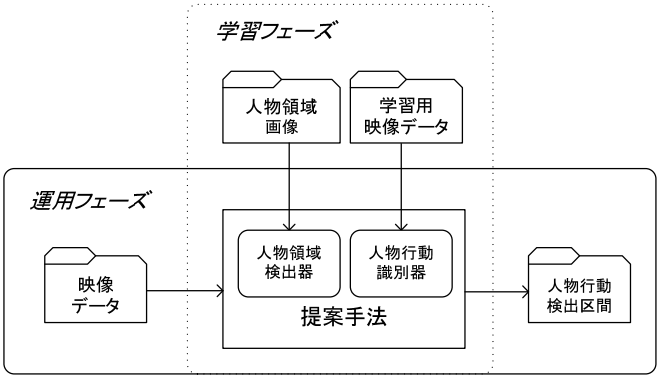


図 3.9　提案手法の概要

提案手法は図 3.10 に示す 4 つのステップで構成される．ステップ 1 では HOG 記述子と SVM 識別器を利用し，映像内から人物領域を検出する．ステップ 2 では 2 次元色ヒストグラムと Kalman フィルタを用いて人物領域を追跡する．ステップ 3 では，人物領域の軌跡特徴と各行動クラスの軌跡特徴を比較し，特定行動を判定する．ステップ 4 では誤検出を抑制するため，判定した特定行動を検証する．次項より各ステップについて述べる．

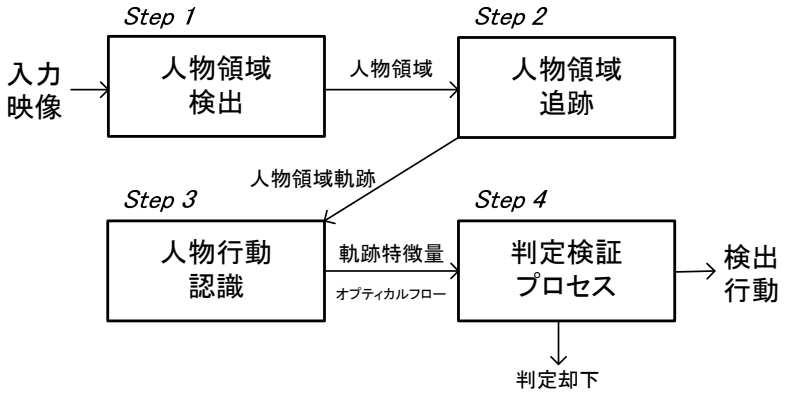


図 3.10　人物行動認識の流れ

○　人物領域検出ステップ

人物領域検出ステップでは，まずフレーム間差分処理を施し，画像内の変化領域を抽出する．人物領域の探索エリアを変化領域に限定することで，人物検出の安定化と処理の高速化を図った．

変化領域内で縦横 2:1 のアスペクト比の矩形（動オブジェクト領域）を位置，サイズを変えながらラスタスキャンし，各領域の HOG 特徴量を算出する．HOG 特徴量は SIFT 特徴量と同様に勾配情報を利用した画像特徴であり，姿勢変化に頑健であるため人物領域検出に広く使用されている．

HOG 特徴量の算出方法は下記の通りである．まず領域サイズの影響を排除するため，動オブジェクト領域を 25×50 pixel に正規化する．正規化した画像を図 3.11 のように“セル”と“ブロック”に分割し，1 セルに付き 9 方向の勾配情報を取得する．1 ブロックでは 81 次元，動オブジェクト領域全体では 1,944 次元（81 次元×24 ブロック）の特徴量となる．

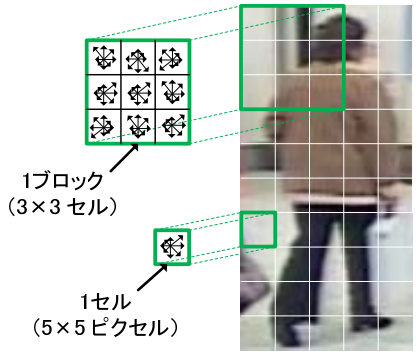


図 3.11　動オブジェクト領域と HOG 特徴量

得られた特徴量を基に，動オブジェクト領域に人物が含まれているか否かを SVM 識別器で判定する．SVM 識別器は，学習フェーズにおいてカメラ毎に 1,000 枚の人物画像（正例）と 2,000 枚の非人物画像（負例）から算出した HOG 特徴量を基に作成した．

識別器から人物領域と判定された領域を人物候補領域とする．これまでの人物検出処理の流れを図 3.12 に示す．

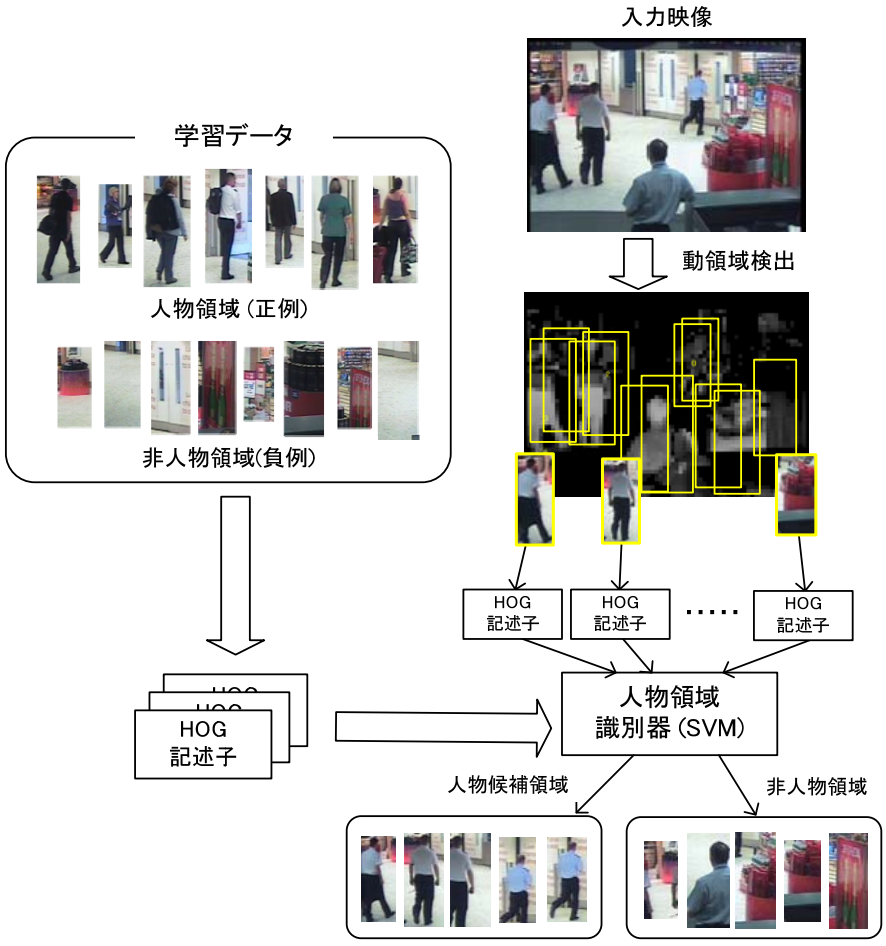


図 3.12　人物検出処理

本手法はラスタスキャンで人物領域を検出しているため，1 人の人物の周辺に複数の人物候補領域が検出される．そこで 1 人につき 1 つの人物領域が割り当てられるよう，人物候補領域のクラスタリングを行った. 各領域の色ヒストグラムと重心間の距離を算出し，類似度が高い領域を同じクラスに割り当てた．各クラスの中心に位置する領域をそのクラスの代表領域として，そのクラスに対応する ID を付与した．

○　人物領域追跡ステップ

隣接フレームの代表領域をマッチングすることで，人物領域追跡を実現する．図 3.13 に追跡処理の流れを示す.

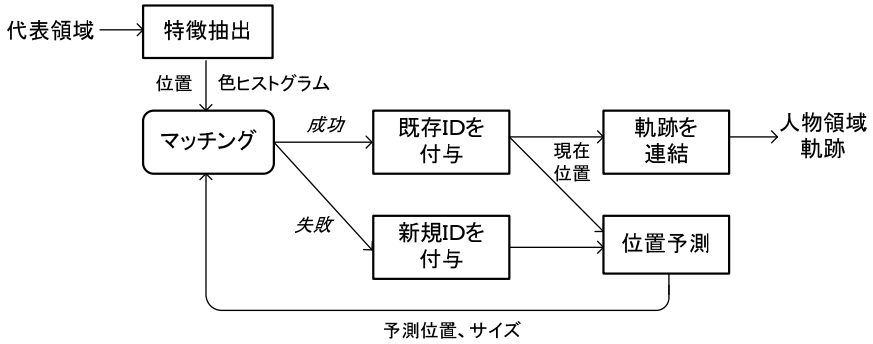


図 3.13　人物追跡処理の流れ

代表領域のマッチングは，人物検出処理のクラスタリング過程と同様，重心間の距離と色ヒストグラムの距離に基づいて行う．まず各代表領域の重心位置を求め，隣接フレームの各代表領域の重心位置とのユークリッド距離を算出する. 次に重心間距離が近い（領域半径未満）代表領域ペアに対して色ヒストグラムを算出する．領域の輝度はその位置の照明条件によって変化するため, 色ヒストグラムには輝度の影響を排除した HSV 色空間の H，S を用いた．ヒストグラム$\boldsymbol{p},\boldsymbol{q}$間の類似度は，式(3-1), (3-2)より算出される Bhattacharyya 距離D_{B}により評価した[Guorong96]．ここで m を成分数とし，ヒストグラム$\boldsymbol{p},\boldsymbol{q}$の総和はそれぞれ 1 に正規化されているものとする．

$$\rho = \sum_{u=0}^{m} \sqrt{\boldsymbol{p}_u \boldsymbol{q}_u} \tag{3-1}$$

$$D_{\mathrm{B}} = \sqrt{1-\rho} \tag{3-2}$$

隣接フレーム間で類似度が高い領域ペアに同じ ID を付与し，マッチング領域が検出されなかった場合は新たな人物領域と判断して新規 ID を付与する．図 3.14 の例に示す

ように, 過去フレームから現在までの代表領域の重心を結合し, 人物領域の軌跡を得る.

人物領域のマッチングには動きベクトルを用いる方法もあるが，混雑した状況では人々が同じ速度で歩くことが多く，各人を動きベクトルで区別することは難しい．色ヒストグラムを用いることで，混雑映像でも比較的頑健に人物領域を追跡できる．

混雑映像では頻繁に人物が交差するため，オクルージョンが多数発生する．そこで Kalman フィルタに基づく人物位置の予測処理を追加した[Grimble94,Yu03]．状態量には画像座標上の人物位置(x, y)および速度(x', y')，観測量には検出した人物領域の重心位置(p_x, p_y)，運動モデルには等速直線運動を仮定した．

追跡中の人物領域の検出に失敗しても，オクルージョンが生じたと仮定し，各フレームでの位置を推定しながら追跡を 1 秒程度継続した．再検出できた場合，検出失敗区間の軌跡を内挿補間するとともに，追跡を再開した．この人物位置予測により，頻繁にオクルージョンが発生する混雑映像においても比較的頑健に人物領域を追跡することができた．

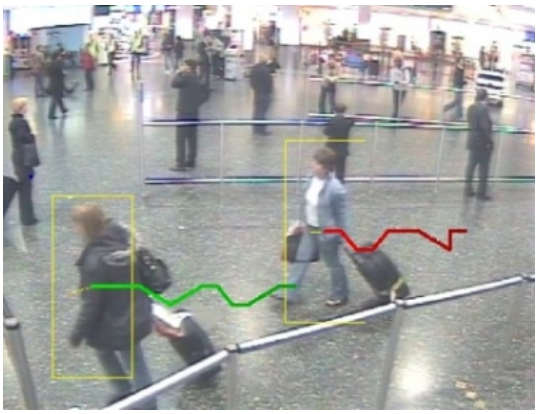


図 3.14　人物領域軌跡の例

○　人物行動認識ステップ

人物の移動軌跡は，速度や方向など，移動を伴う人物行動認識に有効な情報を含んでいる．そこで本手法では特定行動の認識に軌跡から得られる特徴量を利用した．

ただし，軌跡の構成要素である動きベクトルは，画面内の検出位置によりその大きさが異なる．例えば，カメラの近くに位置した人物の動きベクトルは大きく，遠くに位置する人物の動きベクトルは小さい. 本手法では検出位置に関わらず動きベクトルを評価するため，平均速度マップを用いた．

平均速度マップは，入力画像を小さなブロックに分割し，動きベクトル長の平均値をブロック毎に求めたものである．4 時間分蓄積した動きベクトルを学習用データに利用し，各カメラの平均速度マップを作成した．平均速度マップ作成の概念図を図 3.15，平均速度マップの例を図 3.16 に示す. 図 3.16 では矩形内の数値が平均速度を表している．

画像下部から上部へ向かうにつれ平均速度が減少し，人物が全く検出されない最上部のブロックではその値がゼロとなっている．

この平均速度マップを用いて動きベクトルを正規化することで，検出位置に関わらず軌跡特徴量を評価した．さらにこの手法はカメラの位置やその姿勢情報，およびキャリブレーション作業が不要であるため，あらゆる固定カメラ映像に適用できる．

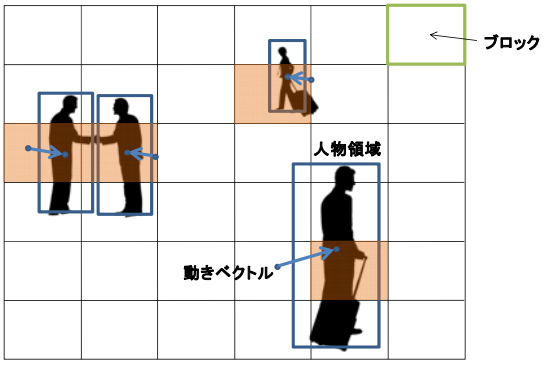


図 3.15　平均速度マップの作成

図 3.16　平均速度マップの例

続いて，軌跡から以下の 7 種（10 次元）の特徴を抽出した．

- 初回検出位置（水平，垂直）
- 最終検出位置（水平，垂直）
- 動きベクトルの総和（水平，垂直）
- 総移動距離
- 平均速度
- 平均加速度
- 直線性

ここで「動きベクトルの総和」は，軌跡内の動きベクトルの合計であり，人物の移動方向を表す．また「移動距離」は初回検出位置から最終検出位置までの距離を表す．この移動距離は動きベクトル同様，平均速度マップを用いて正規化される．次に，「平均速度」は軌跡内の動きベクトル長の平均値を表す．この速度を微分することで「加速度」が得られる．最後に，「直線性」は軌跡内の各位置から軌跡の回帰直線までの平均距離を表す．この値は，人物が直線的に歩いた場合にゼロに近づく．

これらの値は，軌跡が長くなるにつれその特徴が失われる．そこで特徴を抽出する軌跡の区間を 1 秒間に限定した．これは現在の人物検出位置から 1 秒前までの軌跡を扱うことに相当する．図 3.17 に軌跡と特徴量の関係を示す．

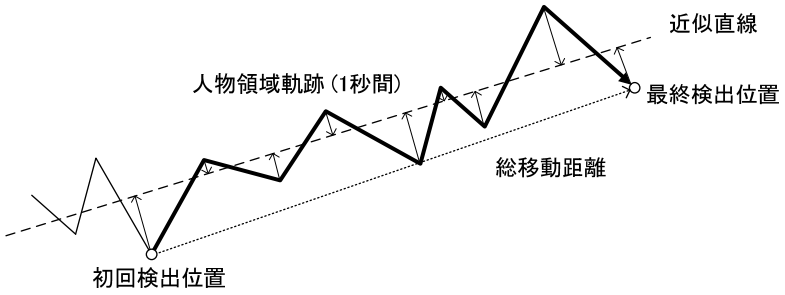


図 3.17　人物領域軌跡と軌跡特徴量

続いて，主成分分析を用いて各軌跡を軌跡特徴空間に投影した．人物追跡により得た軌跡の特徴量から固有値と固有ベクトルを求め，特徴量を 5 次元まで圧縮した．第 1 主成分は「平均速度」と「移動距離」，第 2 主成分は「垂直方向の位置座標」と「動きベクトルの総和」，第 3 主成分は「水平方向の位置座標」，「動きベクトルの総和」，および「加速度」項の比重が顕著であった．

第 3 主成分まで示した軌跡特徴空間上に各軌跡特徴量をプロットしたものを図 3.18 に示す．図中の各点が一つの軌跡特徴を表している．図より，通常の移動軌跡は特徴空間の原点付近に集まり，特殊な移動軌跡は原点から離れた空間に位置する傾向がみられる．さらに，同じ行動の軌跡は特徴空間上で近くに位置する傾向がみられる．

続いて「走る」，「出会う」，「物を置く」の人物行動に着目し，各行動クラスを作成した．図 3.18 における楕円が各行動クラスを表し，その中心$\boldsymbol{\mu}$が平均を，半径が標準偏差の大きさを示す．各行動は，式(3-3)に示す入力軌跡特徴量 $\boldsymbol{x}$ からのマハラノビス距離 $D_M(\boldsymbol{x})$に従って判定した．Σは各クラスの共分散行列を表す．

「走る」クラスは原点から離れた場所に位置するため，軌跡特徴空間上で比較的精度よく識別できる．一方，「物を置く」クラスは原点に近い場所に位置しており，他の軌

跡との分離が難しい．これは「物を置く」行動は移動を伴わないことが多く，移動軌跡からの判定が困難であることを示している．

$$D_M(\boldsymbol{x}) = \sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})} \tag{3-3}$$

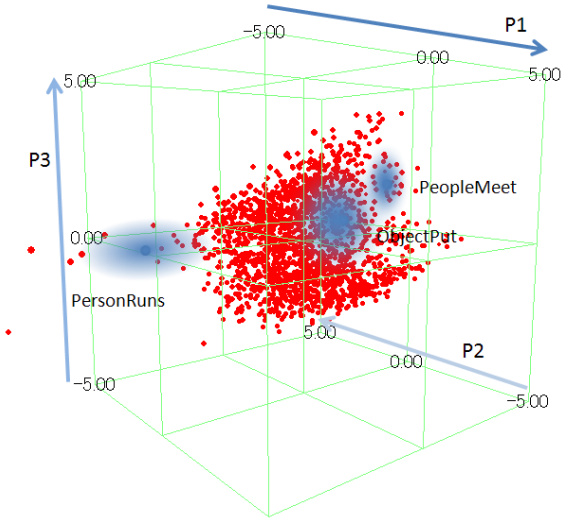


図 3.18　軌跡特徴空間

○　検証ステップ

軌跡特徴空間での判定には誤検出が含まれることが多い．そこで特定行動判定後に検証処理を施した．

「走る」行動は「歩く」行動に比べて動きベクトルが大きいため，ある人物を追跡中に他の人物を誤追跡してしまうことも多い．そこで「走る」行動を検出した後，過去方向へ映像を逆再生し，確実に特定人物を追跡したか否かを検証した．図 3.19 に検証の一例を示す．順方向と過去方向の追跡では人物領域の予測位置が各時点で異なるため，同じ追跡結果となるとは限らない．過去方向の追跡でも「走る」行動が検出された場合のみ，「走る」行動の判定を確定した．

また「出会う」行動においては，行動後に対象人物が大きく移動することは稀である．そこで「出会う」行動判定後も対象人物の追跡を継続し，しきい値 T_d を超える移動量が観測された場合には判定を棄却した．T_d の値は学習フェーズで実験的に定めた．

「物を置く」行動は移動を伴うことが少ないため，移動軌跡だけで判定することは難しい．そこで軌跡特徴空間による検出後，オプティカルフローを評価した．図 3.20 のように下方向への大きなフローが多数検出された場合のみ判定を確定した．

図 3.19　過去方向追跡による検証

図 3.20　オプティカルフローによる検証

3.3.3.　特徴点軌跡ベース(手法 2)

○　手法の概要

次に，手法 2 の特徴点軌跡ベースの人物行動認識手法について述べる．手法 2 は図 3.21 に示す“特徴点検出・追跡”，“特徴量抽出”，“動作識別”の 3 つのステップからなる．

特徴点検出・追跡ステップでは，映像中から多数の特徴点を検出・追跡し，各特徴点の移動軌跡を生成する．続く特徴量抽出ステップでは，各軌跡から固定長の特徴量を作成する．最後に動作識別ステップでは，軌跡記述子を特徴に用いた BoF 法と SVM により人物行動を識別する．BoF 法のコードワードヒストグラム作成する際は，軌跡特徴毎に *tf-idf* 値に基づく重みを考慮した．次項より各処理を詳しく述べる．

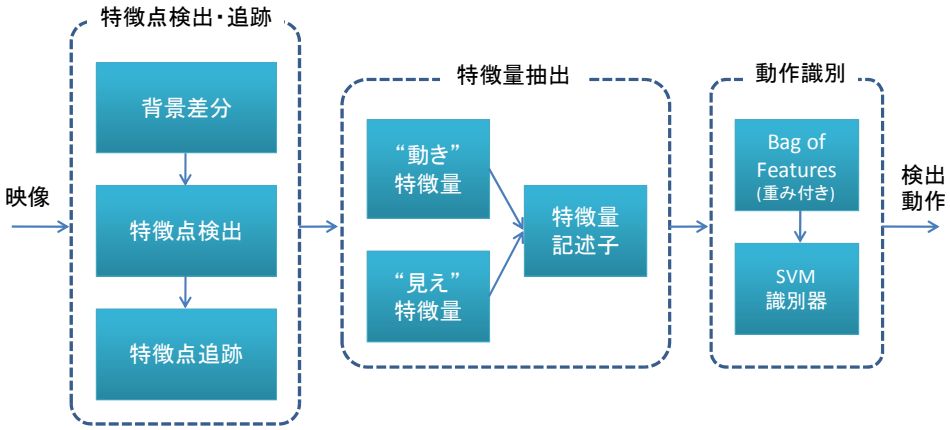


図 3.21　特徴点軌跡ベースの手法の概要

○　特徴点検出・追跡ステップ

▪　背景差分処理

監視映像は固定カメラで撮影されることが多いため，背景差分処理を適用できる．ただし，日射しの変化などによっては輝度変化が生じる場合がある．そこで規定時間における各画素値の平均，分散を求め，背景画像を定期的に更新した．この処理により，人物・荷物などの前景領域と，壁・床などの背景領域を比較的安定して分離できる．

▪　特徴点検出・追跡処理

特徴点検出・追跡処理には Kanade-Lucas-Tomasi (KLT)トラッカを用いた[Shi94]．KLT トラッカは映像解析に広く利用されている特徴点追跡手法である．本手法では特徴点検出に Harris オペレータを用いた．

検出した特徴点を Lucas-Kanade のオプティカルフロー法により追跡した．図 3.22 に追跡中の特徴点軌跡例を示す．前景領域上のみを探索範囲とすることで，人物領域上の特徴点をロバストに追跡した．

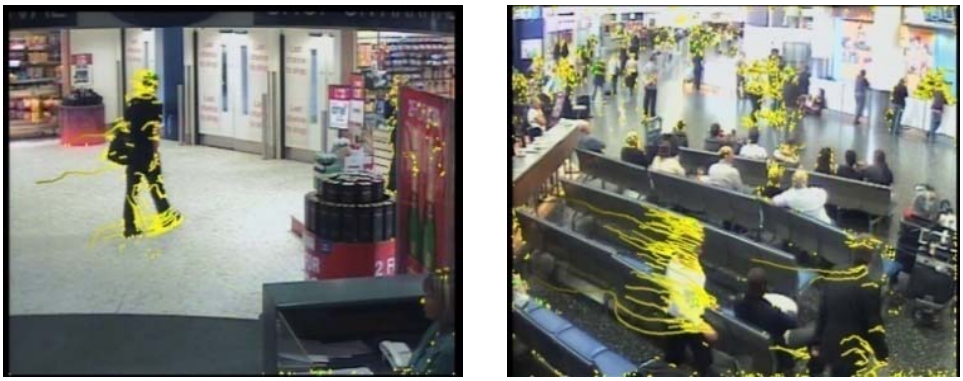


図 3.22　特徴点軌跡の例

○ 特徴量抽出ステップ

▪ “動き”特徴

得られた特徴点軌跡特徴を BoF 法へ入力する．ただし BoF 法は入力に固定次元の特徴量を想定しているため, 時間長の異なる軌跡特徴量をそのまま入力することができない．そこで[Mezaris10]を参考に，各特徴点軌跡から方向に関する固定長ヒストグラム（LIFT 特徴量）を作成した．これを“動き”特徴とする．

下記に LIFT の算出法を示す. まず軌跡の座標データへ Haar ローパスフィルタを施し, 段階的に平滑化する．各段階の軌跡ノードは下記(3-4)-(3-7)式で算出される．

$$\boldsymbol{P}_{u,q} = [\,\boldsymbol{p}^{x}_{u,q}, \boldsymbol{p}^{y}_{u,q}] \tag{3-4}$$

$$\boldsymbol{p}^{x}_{u,q} = \left[\,\boldsymbol{p}^{x,t_1+2^q-1}_{u,q}, \boldsymbol{p}^{x,t_1+2^q}_{u,q}, \boldsymbol{p}^{x,t_1+2^q+1}_{u,q}, \ldots, \boldsymbol{p}^{x,t_2}_{u,q}\right] \tag{3-5}$$

$$\boldsymbol{p}^{y}_{u,q} = \left[\,\boldsymbol{p}^{y,t_1+2^q-1}_{u,q}, \boldsymbol{p}^{y,t_1+2^q}_{u,q}, \boldsymbol{p}^{y,t_1+2^q+1}_{u,q}, \ldots, \boldsymbol{p}^{y,t_2}_{u,q}\right] \tag{3-6}$$

$$\boldsymbol{p}^{x,t}_{u,q} = \frac{1}{2^q}\sum_{i=0}^{2^q-1}\boldsymbol{p}^{x,t-i}_{u,0},\quad \boldsymbol{p}^{y,t}_{u,q} = \frac{1}{2^q}\sum_{i=0}^{2^q-1}\boldsymbol{p}^{y,t-i}_{u,0} \tag{3-7}$$

ここで $\boldsymbol{P}_{u,q}$ を，Q 回を上限に q 段階平滑化した ID 番号 u の軌跡とする．q=0 が平滑化処理前の生軌跡データに相当する．また $\boldsymbol{p}^{x}_{u,q}$ を平滑化軌跡の x 座標集合，t_1, t_2 を検出開始/終了フレームとする．y 座標集合も x 座標と同様の手順で算出する．

各平滑化軌跡中の動きベクトルから, 式(3-8)で示される方向 θ に関するヒストグラムを作成する．この際，R を細分化回数を示す粒度パラメータとし，bin 幅 W_θ を 4 分割, 8 分割，16 分割と段階的に細分化する．各段階でのヒストグラムを総要素数で正規化し，動作時間長に不変な特徴量とした．平滑化段階 q の軌跡の bin 幅 W_θ の方向ヒストグラムを $\boldsymbol{a}(\boldsymbol{P}_{u,q}, W_\theta)$ とすると，式(3-9)で表わされる固定長特徴量 $\boldsymbol{A}_u$ が得られる．

$$\theta^{t}_{u,q} = tan^{-1}\left(\frac{\boldsymbol{p}^{x,t}_{u,q} - \boldsymbol{p}^{x,t-1}_{u,q}}{\boldsymbol{p}^{y,t}_{u,q} - \boldsymbol{p}^{y,t-1}_{u,q}}\right) \tag{3-8}$$

$$\begin{aligned}\boldsymbol{A}_u = \Big[&\boldsymbol{a}\left(\boldsymbol{P}_{u,0}, \frac{\pi}{2}\right), \boldsymbol{a}\left(\boldsymbol{P}_{u,1}, \frac{\pi}{2}\right), \ldots, \boldsymbol{a}\left(\boldsymbol{P}_{u,Q-1}, \frac{\pi}{2}\right)\\ &\boldsymbol{a}\left(\boldsymbol{P}_{u,0}, \frac{\pi}{4}\right), \boldsymbol{a}\left(\boldsymbol{P}_{u,1}, \frac{\pi}{4}\right), \ldots, \boldsymbol{a}\left(\boldsymbol{P}_{u,Q-1}, \frac{\pi}{4}\right) \ldots\\ &\boldsymbol{a}\left(\boldsymbol{P}_{u,0}, \frac{\pi}{2^R}\right), \boldsymbol{a}\left(\boldsymbol{P}_{u,1}, \frac{\pi}{2^R}\right), \ldots, \boldsymbol{a}\left(\boldsymbol{P}_{u,Q-1}, \frac{\pi}{2^R}\right)\Big]\end{aligned} \tag{3-9}$$

図 3.23 に特徴量抽出の概要を示す. 上図の折れ線が特徴点軌跡の生データを示し($\boldsymbol{P}_{u,0}$), 破線が $\boldsymbol{P}_{u,0}$ を一段階平滑化した軌跡($\boldsymbol{P}_{u,1}$)，点線が 2 段階平滑化した軌跡を示す($\boldsymbol{P}_{u,2}$)．本手法では平滑化レベル Q=3，bin 幅粒度レベル R=3 としてヒストグラムを生成した．したがって，一つの軌跡から 84 次元（= 4 bins×3 levels + 8×3 + 16×3）の“動き”特徴量が得られる.

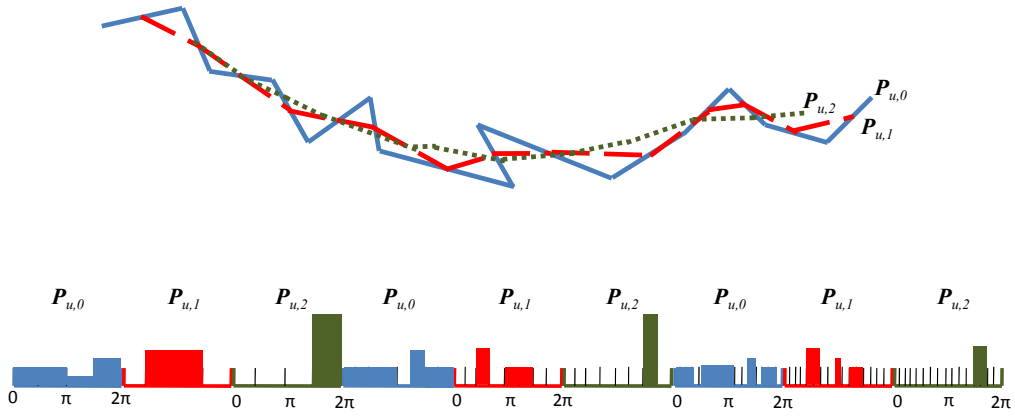


図 3.23　“動き”特徴抽出の概要

- **“見え”特徴**

上記特徴に“見え”特徴を追加することで，“動き”と“見え”の双方を考慮でき，行動認識のロバスト化が期待できる．[Mezaris10]では“見え”特徴として主に SIFT を用いているが，本手法では処理の高速化のため SURF を用いた[Bay08]．各軌跡の追跡終了位置で 128 次元の SURF を記述し，“動き”特徴に付加した．図 3.24 に追跡終了位置までの軌跡例を示す.

最終的に，追跡開始から終了までの 1 本の軌跡につき，212（84+128）次元の“動き”+“見え”特徴量を作成した.

図 3.24　追跡を終了した軌跡例

○　行動識別ステップ

本ステップでは，前項で示した特徴量を基に BoF 法と SVM により人物行動を識別する．BoF 法では規定時間長の映像シーケンスを Bag，212 次元の“動き”と“見え”特徴を Features とみなした．

まず学習フェーズにて，各軌跡特徴を量子化するためのコードブックを作成する．学習用映像から大量の軌跡特徴を収集し，*k*-means アルゴリズムで *k* 個の代表コードワードを作成した．

テストフェーズでは作成したコードブックを参照し，規定時間長シーケンスから抽出した各軌跡特徴を代表コードワードへ量子化する．各コードワードの頻度をカウントし，*k* 個の bin を持つコードワードヒストグラムを作成する．

通常の BoF 法では全特徴を同一の重みで処理するが，この場合，背景領域上のノイズ軌跡の影響を強く受ける．そこで自然言語処理で多く用いられる *tf-idf* 値を利用し，各軌跡特徴の重要度を算出した．

重要度は，式(3-10)に示すように *tf* 値と *idf* 値の積で表わされる．*idf* 値は(3-11)式に示すようにシーケンス頻度の逆数を表し，頻度の高いシーケンスの重要度を下げる働きをする．ここで全シーケンス数を *N*，クラスタ *x* を含むシーケンス数を n_x とする．*tf* 値は式(3-12), (3-13)に示すようにあるクラスタの頻度を表し，特定シーケンス内で主要なクラスタの重要度を高める働きをする．ここで，シーケンス *d* におけるクラスタ *x* の出現数を oc_{xd}，シーケンス *d* におけるクラスタ集合を *W* とした．

したがって *tf-idf* 値を考慮することで，通常のシーケンスに多く現れる背景領域上のノイズ軌跡の重要度を下げ，特定行動シーケンスに多く現れる人物軌跡特徴の重要度を上げる効果が期待できる．

$$Feature\ weight_{xd} = tf_{xd} \times idf_x \tag{3-10}$$

$$idf_x = \log \frac{N}{n_x} \tag{3-11}$$

$$tf_{xd} = \frac{ptf_{xd}}{\sqrt{\sum_{i \in W} ptf_{id}^2}} \tag{3-12}$$

$$ptf_{xd} = 0.5 + 0.5 \times \frac{oc_{xd}}{\max_{i \in W} oc_{id}} \tag{3-13}$$

学習フェーズでは，学習用映像の特定行動シーケンスから軌跡特徴を抽出し，上記手順により重みを考慮したコードワードヒストグラムを作成した．このヒストグラムを最終的な特徴量として正解データ付きの教師付き学習を行い，多クラス SVM 識別器を学

習した．識別行動には「特定行動なし」も含めた．また SVM のカーネルには RBF を用いた．

テストフェーズでは，規定時間長シーケンス内で得られた軌跡特徴からコードワードヒストグラムを作成し，学習フェーズで作成した SVM 識別器でその動作を判定した．シーケンス時間窓を 1 フレーム間隔でスライドしながら，フレーム単位で行動を識別した．その後，各行動の検出区間内での検出頻度に応じて信頼度を定めた．信頼度がしきい値 D を超えた場合にのみ，その行動が生起したと判定した．

3.3.4　特徴点軌跡のクラスタリング（手法 2´）

○　手法の概要

前項では，特徴点軌跡をヒストグラム化し，BoF 法と SVM で識別する人物行動認識手法（手法 2）を紹介した．しかし手法 2 では画面全体を 1 つの Bag として処理したため，人物単位での行動認識が行えないという課題が生じた．

そこで特徴点軌跡ベースの手法に改善を施し，軌跡をクラスタリングして人物単位に近い行動認識を行うこととした．また方向のみで分類していた軌跡ヒストグラムを改良し，方向に加えてオプティカルフローの大きさを考慮したヒストグラムを作成した．さらに高速演算可能な“見え”特徴を用いることで，行動識別処理の高速化を図った．手法 2 の改善版の位置付けであるため，この手法を手法 2′とする．

手法 2′は図 3.25 に示す 4 つのステップで行われる．ステップ 1 では，特徴点の検出と追跡を行う．このステップの処理は手法 2 と同様である．ステップ 2 では，近い位置に存在する軌跡をグルーピングし，画面内の軌跡を複数の集合にクラスタリングする．本研究ではデンドログラム法を用いて軌跡のクラスタリングを行った．手法 2 では BoF 法の Bag に相当するのは画面全体であったが，手法 2´ではクラスタリングした各軌跡集合を Bag として扱った．ステップ 3 では，抽出した軌跡から“動き”特徴と“見え”特徴を抽出し，固定次元の記述子へ変換する．軌跡の時間長に依らず固定次元のヒストグラムへ変換されるため，動作速度の個人差に不変な特徴量となる．最後のステップ 4 では，得られた特徴量記述子から BoF 法と SVM で人物行動を認識する．ステップ 3 で作成した記述子が BoF 法の features として扱われる．事前の学習フェーズにおいて，特定行動中の人物周辺の特徴点軌跡とその行動ラベルを教師付き学習し，SVM 識別器を作成した．具体的には，TRECVID SED データセット内で特定行動を行っている人物のバウンディングボックスを独自に設定し，バウンディングボックス内で抽出した軌跡を正解データとして学習に利用した．

以下では，前項との差分であるステップ 2，ステップ 3 について詳述する．

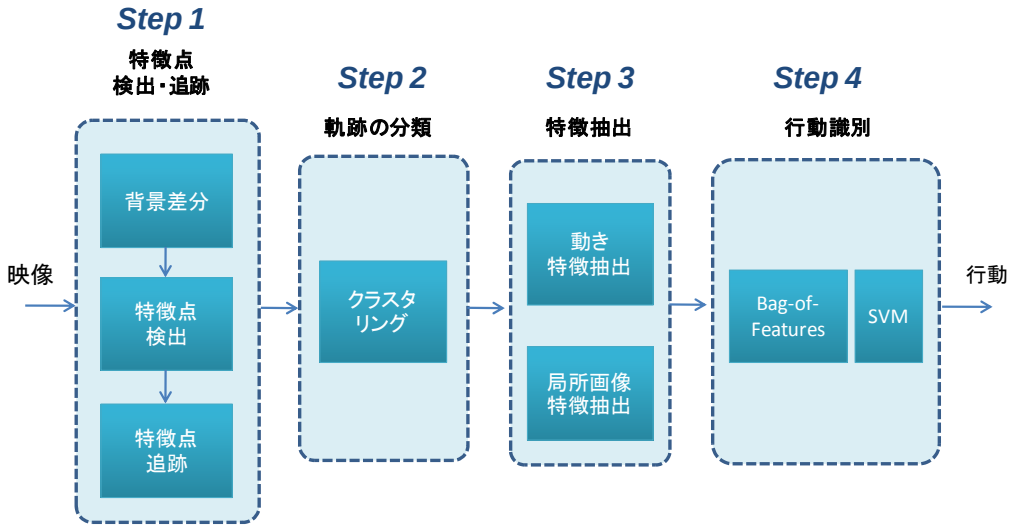


図 3.25　手法 2´ の処理フロー

○　軌跡の分類ステップ

TRECVID SED データセットでは多数の人物が現れ, 多くの特徴点軌跡が生成される. 前章の行動認識処理では Bag を画面全体に設定していたため, 同時に行われている複数の行動を検出することは不可能であった. 本ステップでは, 人物単位の行動認識を実現するため, 抽出した特徴点軌跡をクラスタリングする. クラスタリングした領域を Bag として処理を施すことで, 人物単位の行動認識を目指した.

軌跡のクラスタリングにはデンドログラム法を用いた. デンドログラム法は基本的なクラスタリング手法の一つであり, *k*-means 法のように事前にクラスタ数を与える必要がないことが特徴である. 画面内の人物数は時々刻々と変化するため, クラスタ数を指定する必要のないデンドログラム法は本件に適したクラスタリング手法である.

本研究では軌跡間の距離を評価尺度とし, 近い軌跡同士を一つのクラスタへ統合しながら軌跡のクラスタリングを行った. デンドログラム法ではクラスタ統合処理を打ち切るためのしきい値が必要であるが, 軌跡同士が人物サイズ以上離れた際に統合処理を打ち切ることが望ましい. そこで, 学習フェーズで設定した人物領域のバウンディングボックスの平均サイズを基にしきい値を定めた. これにより人物領域に近いサイズで軌跡集合をクラスタリングでき, 人物単位の行動認識処理が可能となった.

軌跡集合のクラスタリング例は図 3.8 で既に示した. 図中の楕円が人物やオブジェクト領域のおおよそのサイズを表している.

○　特徴量抽出ステップ

▪　“動き”特徴

特徴点軌跡には時間情報が内包されているため，単視点映像から人物動作を解析するための特徴量としては非常に有効である．また，軌跡をフレーム単位で分解すると，各フレームにおける動きベクトルとなる．この動きベクトルを解析することで，特徴点の動きを判断することが可能である．しかしながら，各特徴点の出現から消失までの時間は不定であり，軌跡の長さもその時間に応じて変化する．そのため，固定次元特徴を入力に想定した BoF 法を利用することはできない．そこで本研究では，軌跡を bin 数固定のヒストグラムへ変換した．

手法 2 では動きベクトルの方向のみを考慮したヒストグラムを作成したが，本手法では方向に加え，ベクトルの大きさを考慮した．分割数は方向が 8，大きさが 3 とし，大きさ 0 を含めた 25（= 8*3 + 1）bin のヒストグラムとした．

監視カメラ映像ではカメラに近い人物は大きく，カメラから遠い人物は小さく表示される．そのため，近くの人物の動きベクトルは大きく，遠くの人物の動きベクトルは小さくなる傾向がある．そこで領域クラスタ毎に動きベクトルの大きさを正規化した．

まず，各クラスタ内の軌跡に含まれる動きベクトルの平均長 μ とその標準偏差 σ を算出する．求めた μ と σ に基づき，図 3.26 のようにしきい値を設定して動きベクトルの大きさを，0 を含む 4 段階に分離した．

特徴点の大局的な動きも分析するため，生の軌跡データに加えて 1 段階平滑化した軌跡データも作成した．平滑化には手法 2 同様，(3-4)から(3-7)で表わされる Haar ローパスフィルタを利用した．手法 2´ では Q=1 と設定し，1 段階のみの平滑化軌跡を利用

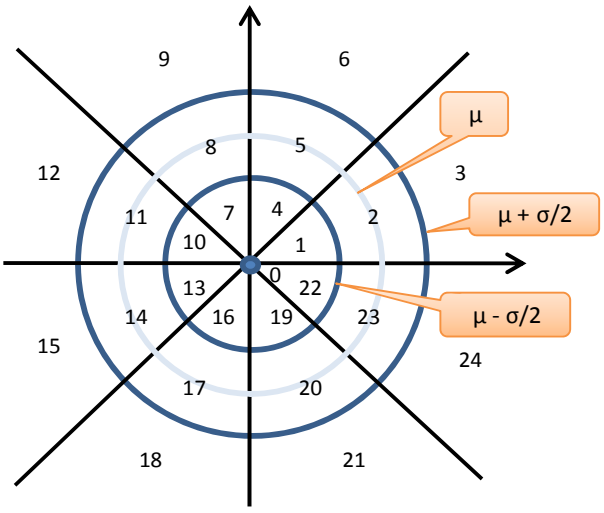


図 3.26　動きベクトルのラベル

動きベクトル長に関するしきい値：　0, (0, μ-σ/2], (μ-σ/2, μ+σ/2], and (μ+σ/2, ∞)

した.

抽出した生の軌跡と平滑化軌跡内に含まれる各動きベクトルにラベルを付与し, 軌跡ヒストグラムを作成した．それぞれ 25bin となるため，合計で 50 次元の“動き”特徴となる.

- **“見え”特徴**

上記特徴は特徴点の動きに関するものであるが，“動き”に加えて“見え”を考慮することで行動認識の頑健化を図る．手法 2 では“見え”特徴に SURF を用いたが，演算コストが高いという問題があった．将来的なリアルタイムアプリケーションを見据え，手法 2´では高速に算出可能な LBP(Local Binary Pattern)を“見え”特徴に用いた[Ojala02].

LBP は注目画素とその周辺画素の画素値の大小パターンを数値化したものである. 本研究では注目画素との隣接 8 画素を周辺画素とし，256 (=2^8)通りの大小パターンを設定した．特徴点を中心とする 16*16 画素領域内の各点で LBP 値を算出し，その頻度ヒストグラムを特徴量として利用した．実際にはノイズの影響を抑制するため，同一軌跡上の各特徴点を中心とした 16*16 画素領域の平均化画像を作成し, その平均画像に関する LBP ヒストグラムを作成した.

最終的に，50bin の軌跡ヒストグラムと 256bin の LBP ヒストグラムを結合し，306 次元の“動き”+“見え”特徴とした.

○　行動識別

人物行動は BoF 法と SVM を用いて識別した．BoF 法では，クラスタリングされた軌跡集合を Bag，306 次元の固定次元特徴量を features として扱った.

まず学習フェーズにて，特徴量を量子化するためのコードブックを作成した．コードブックの作成には *k*-means アルゴリズムを用いた．このコードブックを用いて，クラスタ毎にコードワードの頻度ヒストグラムを作成した．続いて，作成したコードワードヒストグラムとその行動ラベルを利用し，教師付き学習により SVM 識別器を学習した．識別行動には「行動なし」も含めた.

テストフェーズでは，各クラスタで作成した特徴量を基に SVM 識別器でその行動を識別した．行動の識別はフレーム毎に行われるが，各行動の検出頻度により識別の信頼度を算出した. 信頼度が一定のしきい値を超えた場合にのみ, イベント生起を判定した.

3.4　実験

3.4.1　実験条件

上記人物追跡ベースの手法（手法 1），特徴点軌跡ベースの手法（手法 2），特徴点軌跡のクラスタリングの手法（手法 2´）を TRECVID SED データセットに適用し，混雑映像での頑健な行動認識手法を検討した．

それぞれの手法に基づくシステムを試作し，米国 NIST 主催の評価型国際ワークショップ TRECVID SED タスクへの参加を通して性能を評価した. SED タスクでは 5 台の空港監視カメラ 100 時間分の学習用映像と，対象行動が起きた時刻（正解データ）が与えられる．対象行動は「走る」,「物を置く」,「携帯電話をかける」,「指を差す」,「出会う」,「別れる」,「抱擁する」の 7 種類である．また評価用として，正解データのない 50 時間分の評価用映像が与えられる. NIST はこの評価用映像に対する人物行動検出結果をもとに，各チームのシステム性能を評価した．

手法 1 を基にしたシステムでは，「走る」,「出会う」,「物を置く」の 3 種行動を学習した．人物検出に基づく手法では，「走る」など移動を伴う大行動で頑健な検出が期待できる．

手法 2 を基にしたシステムでは，「物を置く」,「指を差す」,「携帯電話をかける」，の 3 種行動を学習した．局所特徴に基づく手法 2 では，比較的小さな動きの検出が期待できる．実験に際しては下記パラメータを用いた．まず Bag の大きさを定める規定時間長を 1 秒（＝25 フーレム）とした．この値は NIST が指定した対象行動の平均動作時間を参考に定めた．また軌跡ヒストグラム作成の際の軌跡平滑化パラメータを Q=3, bin 幅の粒度パラメータを R=3 とした．BoF 法のクラスタ数 k は 1,000 とした．SVM 識別器のカーネルは RBF とし，カメラ毎に識別器を作成した．識別行動クラス数は，上記 3 種の行動に「行動なし」を加えた 4 種類とした．

手法 2´を基にしたシステムでは，「携帯電話をかける」,「抱擁する」,「物を置く」,「指を差す」の 4 種行動を学習した．手法 2 とほぼ同じ識別手法であるが，局所特徴のクラスタリングにより，人物単位の行動認識が期待できる．実験におけるパラメータも手法 2 とほぼ同様である．ただし，手法 2´では領域クラスタを Bag としたため，画面全体が Bag の手法 2 より特徴数が減少する. そこで BoF 法のクラスタ数 k は 100 と設定した．また SVM 識別器は「行動なし」を含む 5 種行動を学習した．

3.4.2　人物追跡ベースと特徴点軌跡ベースの比較

はじめに，手法 1 と手法 2 の精度を比較した．手法 1 システムの対象行動は「走る」,「出会う」,「物を置く」である．一方，手法 2 システムの対象行動は「指を差す」,「携帯電話をかける」,「物を置く」である．「物を置く」行動が唯一共通した特定行動であ

るため，「物を置く」行動の精度で両者を比較した．表 3.1 は NIST から受理した「物を置く」行動の評価結果である．

Detection cost rate (DCR)は，NIST が定めた再現率と適合率を考慮した評価基準であり，値が低いほどシステムの性能が高いことを示す（式(3-14)-(3-17)）．ここで D は信頼度のしきい値，N_{Targ}は観測された特定行動数，$N_{Miss}(D)$はシステムが見逃した特定行動数，$N_{FA}(D)$はシステムが誤検出した特定行動数，T_{Source}はテスト用映像データの合計時間を表す．また本タスクでは式(3-17)に示す定数が用いられた．N_{ref}はテスト用映像に含まれていた「物を置く」行動数，N_{sys}は各年度のシステムが検出した「物を置く」行動数，N_{CorDet}は各年度のシステムの正解数である．

表 3.1 によると，手法 1 システムの正解数 19 に対して手法 2 システムの正解数は 39 であり，「物を置く」行動に関しては手法 2 の精度が高いことが分かる．この結果は，単一軌跡より多数の軌跡を考慮することで，行動識別の性能が向上することを示している．特に，TRECVID データセットのような非常に混雑した映像ではオクルージョンが

$$DCR(D) = P_{Miss}(D) + \beta * R_{FA}(D) \tag{3-14}$$

$$P_{Miss}(D) = \frac{N_{Miss}(D)}{N_{Targ}}, \quad R_{FA}(D) = \frac{N_{FA}(D)}{T_{Source}} \tag{3-15}$$

$$\beta = \frac{Cost_{FA}}{Cost_{Miss} * R_{Target}} \tag{3-16}$$

$$Cost_{Miss} = 10, \ Cost_{FA} = 1, \ R_{Target} = 20 \tag{3-17}$$

$N_{Miss}(D)$: *The number of missed detections at decision score* D

N_{Targ} : *The number of event observations*

$N_{FA}(D)$: *The number of false alarms at decision score* D

T_{Source} : *The total duration of the video segments in hours*

表 3.1　「物を置く」行動での手法 1 と手法 2 の比較

手法	N_{ref}	N_{Sys}	N_{CorDet}	N_{FA}	N_{Miss}	R_{FA}	P_{Miss}	DCR
人物検出ベース	621	488	19	469	602	30.760	0.969	1.123
特徴点軌跡ベース	621	1061	39	537	582	35.219	0.937	1.113

大量発生するため，人物領域を確実に検出することが難しい．手法 1 では精度が人物検出精度に大きく依存したため，混雑映像では頑健な処理ができなかったものと考えられる．これらのことから，混雑した実映像に対しては，人物追跡ベースの認識より多数の特徴点軌跡ベースの手法が適していると考えられる．

表 3.2 は NIST から受理した結果をもとに，独自に適合率，再現率を算出した結果である．手法 1，2 システムで識別した全行動に対する結果を示している．共通行動の「物を置く」では，手法 1 が再現率 3.06%，手法 2 が 6.28%であった．再現率で評価してもやはり手法 2 で高い精度が得られた．

一方，手法 1 では「走る」行動で再現率 14.02%と高い精度を見せている．また「出会う」行動に対しても 12.25%と比較的良好な結果を示している．この結果は，人物の移動を伴う大きな行動に対する，手法 1 の有効性を示している．

手法 1 で「走る」行動の精度に貢献したのは，HOG と SVM ベースの高精度な人物検出処理と，Kalman フィルタベースの頑健な人物追跡処理であった．上記 2 処理により，対象人物の軌跡を正確に取得できたことが，「走る」行動の見逃し数を減らし，再現率の向上につながったと考えられる．「走る」クラスが軌跡特徴空間上で第一主成分に大きな偏りがあったことも検出精度を高める要因となった．さらに過去方向追跡に基づく検証プロセスも，適合率向上に大きく貢献した．

ただし，突発的な駆け出しなど，短期的な「走る」行動を検出することは困難であった．これは軌跡の最長区間を 1 秒に限定して処理したためである．したがって，軌跡の最長区間は検出対象に合わせて調整する必要があると考える．

表 3.2　手法 1 と手法 2 の比較

人物追跡ベース

行動	リファレンス数	検出数	正解数	誤検出数	見逃し数	適合率 (%)	再現率 (%)
走る	107	354	15	339	92	4.24	14.02
出会う	449	960	55	905	394	5.73	12.25
物を置く	621	488	19	469	602	3.89	3.06

特徴点軌跡ベース

行動	リファレンス数	検出数	正解数	誤検出数	見逃し数	適合率(%)	再現率(%)
物を置く	621	1061	39	537	582	3.68	6.28
指を差す	1063	3495	54	884	1009	1.55	5.08
携帯電話をかける	194	52	1	41	193	1.92	0.52

また「出会う」行動クラスは，軌跡特徴空間で位置と加速度の影響が顕著な第三主成分で偏りが見られた．これは，空港では立ち止まる位置がおおよそ限定されており，また「出会う」際に動作者の移動速度が急激に落ちるためと考えられる．

ただし，今回の手法では単一軌跡のみを対象として検出処理を行ったため，一人の人物が立ち止まっただけでも「出会う」と誤検出される例が見られた．このような複数人数での行動に関しては複数の軌跡を考慮するなど，改善を検討する必要がある．

人物の移動を伴わない小行動の「物を置く」は，適合率，再現率ともに低い値となった．軌跡特徴空間ではクラス中心が原点付近に位置したため，他の行動との識別が困難であった．また「物を置く」行動の定義が広く，「テーブルに物を載せる」，「カートを立て掛ける」なども同行動に属された．そのため，単一方向のみを考慮したオプティカルフロー解析では識別が困難であった．

図 3.27 は，各手法で対象とした行動の TRECVID SED タスク参加チーム内順位を示したものである．この順位は，NIST から受理した DCR 値を他の参加チームと比較して定めた．この表より，手法 1 は大行動である「走る」の検出に優れているものの，他の行動の検出精度は低く，汎用性が低いことが分かる．一方，軌跡ヒストグラムに基づく手法 2，手法 2´では，特定行動で突出した精度は見られないものの，どの行動に対しても比較的高い性能を見せている．汎用性を考慮した場合，局所特徴に基づく人物行動認識が優れていることが分かる．

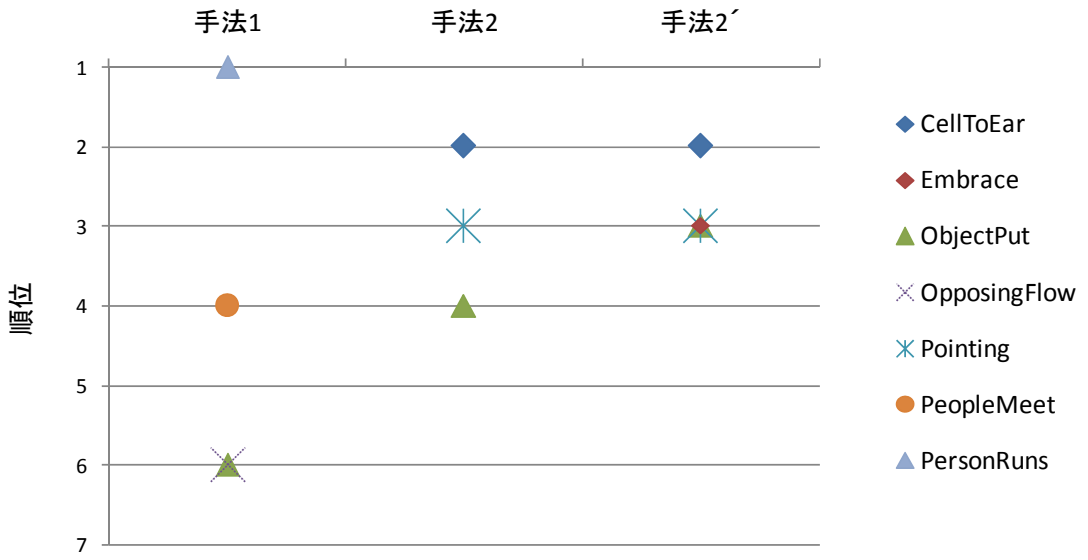


図 3.27　人物行動毎の各手法の順位

3.4.3　軌跡のクラスタリング手法の検証

本項では手法 2´である軌跡のクラスタリング手法について検証する．NIST からの評価結果を基に算出した手法 2´の適合率，再現率を表 3.3 に示す．前項の表 3.2 と比較すると，共通行動「物を置く」で精度が悪化していることが分かる．手法 2´は手法 2 の改良版としての位置付けであったが，良好な結果は得られなかった．その理由を以下に考察する．

図 3.28 は軌跡のクラスタリングを施した一例である．図中の楕円が一つのクラスタを示している．この図より，クラスタ内の軌跡数が非常に少ないことが分かる．Bag 内の特徴サンプル数が少ない場合は，そのコードワードヒストグラムがスパースとなり，精度の高い識別が困難となる. 画面中の全軌跡を一つの Bag で処理した手法 2 では十分な軌跡数を確保できるため, このような軌跡ヒストグラムのスパース化は起こりにくい.

手法2, 手法2´で作成した軌跡ヒストグラムからそれぞれ30個をランダムに選択し，bin 数に対する非ゼロ要素の割合を表 3.4 に示した．ヒストグラムの bin 数(=コードワード数 k) は手法 2 が 1,000，手法 2´は 100 である．手法 2´の bin 数を少なめに設定したにも関わらず, 手法 2´では非ゼロ要素の割合が 7.5%と, 非常にスパースであった. 画面から抽出する特徴点の数を増やすなど, 軌跡ヒストグラムのスパース化を防ぐ手段を講じるべきであった.

また図 3.28 において，画面手前の人物に対しては精度の高い領域クラスタリングができているものの，画面奥の群衆に対しては 5 名程を 1 つのクラスタに含めている．これは，デンドログラム法のしきい値を固定していたことに起因する．今回解析対象としたような俯瞰映像においては，画面手前の人物が大きく，画面奥の人物が小さく映る．そのため位置を考慮し, クラスタリングしきい値を動的に定める必要があったと考えられる．

さらに，今回のクラスタリング処理は各フレーム毎に行っており，隣接フレームや時間情報を考慮していない．そのため，突発的に出現するノイズクラスタの影響を避けられず，False Alarm の大量発生を招いてしまったとも考えられる．

表 3.3　軌跡のクラスタリング手法の行動認識精度

行動	リファレンス数	検出数	正解数	誤検出数	見逃し数	適合率 (%)	再現率 (%)
携帯電話	194	1447	3	162	191	0.21	1.55
抱き合う	175	3869	31	804	144	0.80	17.71
物を置く	621	9216	10	552	611	0.11	1.61
指を差す	621	13974	41	1237	1022	0.29	6.60

上記のような課題を改善することで, 軌跡のクラスタリングによる手法の精度は向上すると考えられる．今後のシステム設計の際には十分に検討する必要がある．

図 3.28　軌跡のクラスタリングの例

表 3.4　軌跡ヒストグラムの非ゼロ要素の割合

	手法2 （非クラスタリング）	手法2´ （クラスタリング）
非ゼロ要素の割合(%)	19.5	7.5

3.4.4　特徴点軌跡ベースの手法の検証

本項では，本研究で最も高い性能を示した特徴点軌跡ベースの手法（手法 2）に対する検証を行う．

○　従来手法との比較

まず，TRECVID SED 開発用データセットを用いて提案手法を従来手法の Trajectons 法[Matikainen09]と比較した．Trajentons 特徴 $\boldsymbol{T}$ は下式で与えられる．ここで特徴点の位置(x, y)を $\boldsymbol{X}$，特徴点の ID 番号を i，またフレーム番号を t とする．各軌跡は時間長 L の軌跡片に分割される．今回の実験では，L=10 [frame]と設定した．

$$\boldsymbol{T}_i^t = \left\{\boldsymbol{X}_i^t - \boldsymbol{X}_i^{t-1}, \boldsymbol{X}_i^{t-1} - \boldsymbol{X}_i^{t-2}, \ldots, \boldsymbol{X}_i^{t-L+1} - \boldsymbol{X}_i^{t-L}\right\} \tag{3-18}$$

提案手法では，*tf-idf* 法で重み付けされた LIFT 特徴記述子を特徴量に用いて識別器を作成した．それぞれの識別器は，TRECVID 開発用データセットの中のカメラ 1 の映像 8 時間分を用いて学習した．

TRECVID 開発用データセットの中で，学習に用いた部分以外の 2 時間分のデータをテスト用映像とし，その中で 111 回発生した「指を差す」動作で評価した．監視カメラ映像解析においては，適合率よりも再現率を重要視すべきであるため，本実験では再現率を評価指標に用いた．

Trajectons 法による再現率は 15.32%，提案手法による再現率は 18.92%であった．この結果は，従来の特徴点軌跡解析手法に対する提案手法の有効性を示している．Trajectons 手法で扱う軌跡は，タイムワーピング法などを用いておらず，その時間長が L に固定される．そのため，同じ行動でも速度が異なる場合には異質の特徴量とみなされることが誤認識の原因と考えられる．さらに，L より長い行動を検出することは難しい．このように従来手法では，固定長の特徴量を用いることにより情報の欠落が生じていると考えられる．

○　“動き”特徴，“見え”特徴の検証

本実験では SURF と LIFT の 2 種類を従来手法に用いた．SURF は“見え”特徴のみを有し，LIFT は“見え”と“動き”特徴を有する．提案手法は LIFT 特徴に加え，各特徴の重要度を考慮した手法である．TRECVID SED 開発用データセット 40 時間分の開発用映像で識別器を学習し，他の 10 時間分の開発用映像で評価した．

表 3.5 に各手法の精度を比較した結果を示す．#Ref 列内の数値は評価用映像内の特定行動数を示し，SURF，LIFT，提案手法 列内の数値は各行動に対する精度を示す．監視システムの評価においては一般に適合率より再現率を重視するため，精度の評価指標として 5 台のカメラ映像の平均再現率を用いた．

SURF と LIFT の比較結果より，“動き”特徴が特定行動の識別精度を飛躍的に向上させていることが分かる．この結果は，人物行動認識では“見え”特徴だけでは不十分であり，“動き”特徴が不可欠であることを示している．

提案手法では，LIFT よりさらに高い検出精度が得られた．これにより，画像特徴が微小となる局所行動に対しては，特徴の重みが有効に機能することが示された．

「携帯電話をかける」行動はいずれの手法も検出に失敗した．これは映像中の画像特徴が非常に微小であり，検出に有効な軌跡が十分に得られなかったことが原因と考えられる．軌跡の検出位置を考慮するなど，さらに有効な特徴量を今後検討する必要がある．

「特定行動なし」の再現率は提案手法が最も低い結果となった．これは「特定行動なし」シーケンスを何らかの特定行動と誤認識したためである．誤検出率は増えたものの，

その他の特定行動の検出精度は向上しているため，システム全体としての性能は提案手法が優れていると考える．

表 3.5　SURF，LIFT との比較

行動	#Ref	SURF (%)	LIFT (%)	提案手法 (%)
指を差す	276	3.99	6.16	8.33
物を置く	130	0.77	9.23	11.54
抱擁する	76	2.63	25.00	28.95
携帯電話	74	0.00	0.00	0.00
特定行動なし	667	97.30	94.15	92.80

再現率(%)

○　TRECVID SEDタスクによる評価

手法 2 に関し，NIST から受理した TRECVID SED タスクの結果を表 3.6 に示す．検出対象行動は「物を置く」，「指を差す」，「携帯電話をかける」である．ここで，#Ref はテスト用映像内の特定行動数，#Sys は本システムが検出した特定行動数，#CorDet は本システムの正解数を示す．

図 3.29 に評価結果の Detection error tradeoff (DET)カーブを示した．システムは信頼度がしきい値 D を超えた時点で特定行動の生起を判定するが，D の値を変化させながら DCR をプロットしたグラフが DET カーブである．横軸が誤検出率，縦軸が見逃し率に相当し，システムの性能が高いほどグラフは原点に近付く．

「物を置く」行動は他の行動に比べて高い精度が得られた．これは「置く」行動では下向きの同一方向ベクトルが現れやすいためと考えられる．一方，「指を差す」行動は右方向，左方向など動作方向のバリエーションが広いため，識別が困難であったと考える．今後，特定行動を方向別に学習するなど，学習方法を見直す必要がある．

「携帯電話をかける」行動は比較的高い DCR が得られたものの，検出に成功したのは 1 シーケンスのみであり，満足できる性能とは言えない．微小行動の検出に有効な特徴量をさらに検討する必要がある．

表3.7に手法2を他チームと比較した結果を示す．手法2システムのDCRは全ての行動において平均DCR値を上回っており，提案手法の効果を確認できた．

実応用化に向けてはまだまだ課題も多いが，軌跡特徴の位置情報や共起性を考慮するなど様々な改善を施し，さらなる性能向上を検討したい．

表 3.6　TRECVID SED タスクの結果

行動	#Ref	#Sys	#CorDet	DCR
指を差す	1063	3495	54	1.239
携帯電話	194	52	1	1.008
物を置く	621	1061	39	1.113

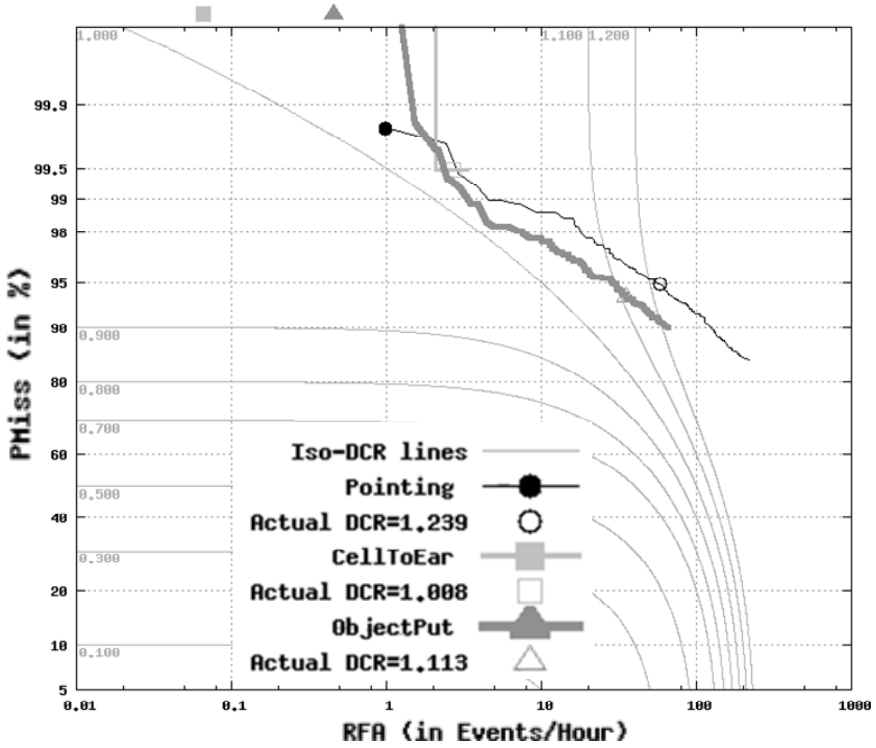


図 3.29　DET カーブ

表3.7 他システムとの比較

行動	DCR（提案手法）	DCR（平均）	標準偏差
指を差す	1.239	5.400	5.760
携帯電話	1.008	3.240	2.514
物を置く	1.113	2.665	1.932

3.5　まとめ

本章では混雑した映像から特定行動を頑健に検出する手法を検討した. 主に人物追跡ベースの手法と軌跡特徴量ベースの手法の比較を通し, 混雑した映像での頑健な人物行動認識手法を模索した．人物追跡ベースの手法では，広域特徴に基づく人物検出により人物の移動軌跡を作成し，その単一軌跡から人物の行動を認識した．人物の多様な外見に対応するため，HOG 特徴を用いて人物領域を検出した．軌跡特徴量ベースの手法では，多数の局所特徴点軌跡を Bag-of-features 法に適用して人物行動を認識した．激しいオクルージョンに対応するため，あえて人物非検出の手法を用いた.

人物追跡ベースの手法は「走る」などの特定の大行動を頑健に検出できる一方，その精度は人物検出精度に依存した. 軌跡特徴量ベースの手法は人物非検出のため多様な外見やオクルージョンに頑健であり，突出した精度は見せないものの，多種の行動を比較的頑健に検出した．軌跡特徴量ベースの手法は，混雑映像中でも「物を置く」行動を再現率 6.28%で検出した．これら 2 手法の比較により，混雑映像に対する局所特徴点軌跡ベースの手法の優位性を確認した.

またあらゆる時間長の軌跡特徴量を固定次元ヒストグラムへ変換することで, 動作速度の個人差に不変な特徴量を作成した. 従来手法との比較や TRECVID SED タスクへの参加を通し，本研究の提案する特徴量の有効性を確認した.

さらに特徴点軌跡ベースの手法を発展させ, 軌跡のクラスタリングベースの手法を検討した．特徴点軌跡ベースでは画像全体で 1 つの行動を検出したが，軌跡をクラスタリングすることで人物単位の行動認識処理を目指した. 評価結果は予想よりも低いものであったが，詳細な分析により，今後の精度向上へ向けた方向性を定めた.

本研究で得られた知見は, 動作解析に関する幅広いアプリケーションに応用可能なものである．たとえば続く 4 章においては，本章で考案した軌跡ヒストグラムを用いて人物の情動動作の検出を行っている．今後も本研究で培った技術を様々な分野の研究・開発に活かしていきたい.

第4章　情動計測による注目度推定

4.1　はじめに

近年，人間観測による内部状態推定技術への需要が高まっている．脳波をはじめ，カメラ映像などの可視情報や専用機器で測定した視線変動，生体反応などを利用し，被験者の集中度や感情などの内部状態を推定する研究が行われている[Kagaya05, Umemoto11, Sase10].

たとえば映像コンテンツを視聴しているユーザの注目度を推定できれば, そのユーザの趣味・嗜好を理解でき，適切な映像コンテンツを推薦することが可能となる．また注目度を活用することで，“視聴率”に代わる“視聴質”の計測が可能になると期待されている. しかし, 不可視である人物の内部状態を計測することは本来非常に困難である.

現在のところ，脳波や心拍数，発汗量，視線変動量などの生体信号を特徴量とし，人物の注目度や集中度が高まるメカニズムを科学的に探究・解明するための研究がなされている. ただし, このような生体信号計測は一般に接触型の専用デバイスが必要となり，視聴者の視野や身体動作の自由度を奪うため，家庭での利用には向いていない.

一方で，家庭での映像コンテンツ推薦などのアプリケーションを想定し，工学的アプローチに基づいて人物の内部状態を推定する手法も研究され始めている. 例えば非接触型のカメラを入力とし, 表情をもとに感情推定を行っている研究も多い. 感情推定には，視聴者の表情識別が有効と考えられる．ただし，映像解析に基づく人物の感情推定技術は未だ確立されておらず，十分な信頼性があるとは言えない．また視聴者感情の表出には個人差があり，表情に感情が殆ど表れないユーザも多い．特に今回ターゲットとしたコンテンツへの注目状態は，視聴者の表情に表れにくいと考えられる.

心理学の分野では，人物の内部状態を判断する材料として，表情とともに身体動作が検討されてきた[Handa01]．工藤は，人は欺瞞場面において感情を取り繕う際には顔の表情を意図的に操作しようとする傾向があるため, 真の感情状態は無意識的に身体動作に反映されやすいと述べている[Kudo99]．また Morris も，顔面表情は最も自己認識が進んだ部分であることから意図的に抑制することができる一方で, 身体の中でも普段隠蔽されている部分には真の感情状態が漏えいしやすいと述べている[Morris91]．さらに Argyle も同様に，身体動作は顔には表れない感情を伝達するものであり，抑制された言語や顔には表れなかった感情状態が漏えいすると述べている[Argyle75].

以上の先行研究から，身体動作は人物の内部状態推定の有効な指標となることが予想される．序章で述べたように，人物動作は意図的動作と無意図的な情動に分けられ，さらに情動は筋運動系，自律神経系，内分泌系・免疫系に分類される．筋運動系情動は人間の感情と密接に関係する無意識の身体動作であり，カメラや非接触型センサで計測可能である．本研究では，汎用機器である Microsoft Kinect を用いて筋運動系情動を計測することとした．

Kinect とは Microsoft のゲーム機 Xbox の姿勢計測センサである [Microsoft10]．Kinect はセンサの前に立つユーザの頭部，腕，足などの詳細な身体部位の 3 次元位置をリアルタイムで計測できる．具体的には無数の赤外線パターンを照射し，赤外線カメラで被写体に照射されたパターンを観測する．パターンの投影位置から，被写体までの奥行きを計測し，人物領域を検出する．その後，事前に学習した膨大な人物姿勢パターンと観測した人物領域を比較し，各身体部位の 3 次元位置を正確に推定する．

Kinect はリモコンなしのジェスチャ操作でゲーム機を制御するセンサとして開発された．しかし近年ではその性能の高さから研究用途にも積極的に活用されている．また指先領域認識や表情認識など，今後の発展も期待されるデバイスである．図 4.1 に Kinect を用いた身体部位の計測例を示す．各部位の 3 次元位置を 2 次元画像平面へ投影することで，身体部位を 2 次元画像座標で計測することも可能である．

また，非接触デバイスで内部状態を推定するための特徴として，多くの研究で“視線”が用いられている．「目は心を映す鏡」とも言われており，視線は内部状態を探るための重要な特徴となり得る．しかし，家庭用汎用機器で精度良く視線を計測することは難しい．そのため，視線は専用の視線計測器を用いて計測されることが多い．視線計測器には，ゴーグル型およびキャップ型の接触型デバイス，もしくは非接触型のアイトラッキング用専用カメラなどがある[NAC]．それぞれの例を図 4.2，4.3 に示す．双方とも視線計測の原理は同じであり，黒目領域と瞳に映った光彩点の情報を基に視線を計測している．

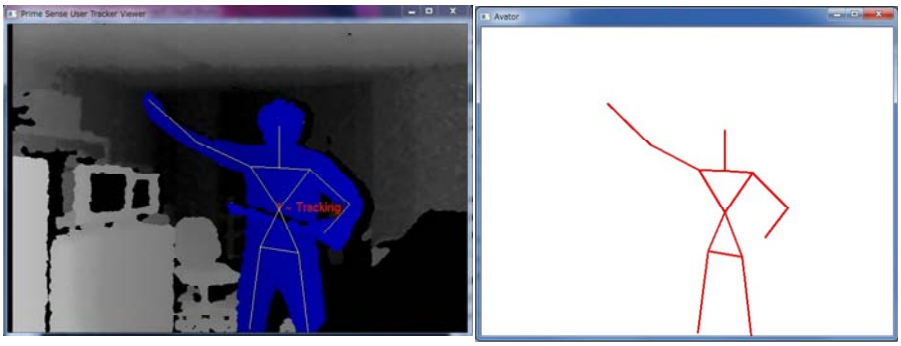


図 4.1　Kinect での身体部位位置計測例

これらのデバイスを使用することで, 非常に精度の高い視線位置を計測することができる．しかし接触型視線計測器はデバイスを身体に装着する必要があり，家庭内での使用には適していない．また非接触型のアイトラッキング用カメラも，頭部をある程度固定して計測する必要があり日常生活での利用は難しい．

そこで, 家庭内で利用可能な汎用機器を用いて映像コンテンツへの注目度を推定する手法を検討した．近年のカメラは一般商用モデルにおいても高解像度化が進み，画面の細部における特徴も計測可能となっている．またＴＶ受像機にはカメラを搭載したモデルも既に発売されており, 近い将来は映像解析による視聴者の状態推定が一般化する可能性が高い．このような視聴者カメラで視聴者の瞳領域の光彩点を抽出することは解像度的に困難であるが，黒目領域を抽出し，その動きを解析することは十分可能である．そこで，通常の HD カメラも入力デバイスに用いた．

筋運動系情動のような微小動作はその人物の内部状態を映し出す一方で, 外的要因の影響も受けやすい．たとえば，動物は動くものを目で追う習性があるため，本人の興味や注目度に関わらず視線が変動することも多い．このように，外的要因の影響を考慮することも内部状態推定には必要と考えられる．

図 4.2　接触型視線計測器

図 4.3　非接触型視線計測器

以上を鑑み，本研究では下記の課題に取り組んだ.

1. 映像解析による可視情報からの人物の内部状態推定
2. 家庭用汎用映像機器での無意図的動作（情動）計測
3. 内的要因のみならず，外的要因の影響を考慮した内部状態推定

非接触型センサから人物の内部状態を推定することは本来困難である. 本研究では内部状態の推定に有効な特徴量を検討するとともに，その可能性を検証する．また家庭での実運用を目指し，使用するセンサを家庭用汎用機器に限定する．さらに筋運動系情動に影響を与える外的要因を検討し, 内的要因と外的要因双方を考慮した内部状態推定器を提案する.

4.2　関連研究

4.2.1　履歴に基づく推薦

インターネットの普及により，各家庭における電子商取引が一般的となっている．この電子商取引はＰＣを介して行われることが多く, 個人の購買履歴や操作ログを利用して新たな商品の購入を促すサービスが生まれている. この技術をＴＶ番組推薦へ応用し, 個人の視聴履歴と EPG データなどの番組メタデータからユーザの趣味・嗜好を理解する研究が行われている[Igawa10].

また商品の購買履歴以外にも, 閲覧した Web ページの履歴や他のユーザの購買履歴を集合知的に利用した推薦サービスも存在する. 特に後者は協調フィルタリングと呼ばれ, 購買データベースから消費者の購買傾向を理解し，関連商品を的確に推薦できるため, 多くの電子商取引サイトで用いられている. これらの技術をＴＶコンテンツの推薦に応用した例もある[Hijikata07, Yamaguchi10].

これら個人や他者の視聴履歴からユーザの趣味・嗜好を理解し，ＴＶコンテンツを推薦する手法は, 誤差となるノイズデータが少なく, 確度の高い推薦を行うことができる. またカメラなどのセンサが不要であり，実環境での利用に適している．しかし個人プロフィールの登録など能動的行動を要する場合が多く, 子供や高齢者など複雑な操作が不得手なユーザにとってはサービス利用の障壁となるおそれがある. またこれらの手法におけるユーザ理解は番組単位で行われることが多く, シーン単位での粒度の細かいユーザの趣味・嗜好の理解を実現した例は少ない．そのため，シーン単位でのコンテンツ推薦は未だ困難である.

さらに, 視聴履歴に基づいた手法では推薦する番組ジャンルが固定化する傾向があり, ユーザの潜在的な需要に基づく意外な番組が推薦されることは期待できない. 視聴履歴

のようなユーザの能動的行動情報のみならず，各種センサで観測したユーザの非能動的行動情報を用いてユーザを理解することが必要である．

4.2.2　身体特徴計測による内部状態推定

ユーザの非能動的行動情報を理解する手段として，脳波や心拍数，発汗量，瞳孔径などの生体信号を利用する方法がある．梶原は，脳波からドライビング時の負担を計測した[Kajiwara11]．この研究ではアルファ波で集中度，ベータ波で安らぎ度を観測し，運転負荷が生じると集中度が上昇し，逆に安らぎ度は低下することを確認した．このように集中度などの内部状態推定に対しては脳波計測の有効性が確認されている．しかし正確な脳波計測には多数の電極を頭部に装着する必要があり，日常的なＴＶ視聴行動での利用には適さない．

脳波と並び，ユーザの内部状態を推定するための特徴量に注目されているのが視線である[Yasuma11, Just76, Emery00, Salvucci01]．視線は脳波ほど大規模な計測器を用いずに計測できるため，視線からユーザの集中度や番組内容の理解度を推定する手法が数多く行われている．米谷らは，視線運動と映像コンテンツの顕著性変動の双方を考慮し，高い精度でユーザの集中度を推定する手法を提案した[Yoneya11]．ただし正確な視線方向の計測には接触型の視線計測器を用いる必要があり，これも日常的な家庭での利用には適さない．

そのため，映像から集中度などの視聴者の内部状態を推定する試みが行われ始めている．山北らは，ドライビング時の顔表情変化から集中状態を推定する手法を提案した[Yamakita06]．この手法ではユーザの顔を撮影した映像から視線，口の動き，瞬き回数を計測し，集中度の推定に利用した．このような非接触型デバイスを用いた手法はユーザへの負担も少なく，家庭での利用に適している．

ただし，映像解析に基づくユーザ特徴の計測は一般に精度が低く，内部状態の推定に大きな誤差を伴うことが多い．そのため，精度の高い特徴計測器の作成が求められている．加えて，ユーザの内部状態推定に有効に寄与する観測可能な身体特徴は未だ十分に検討されていない．視線や瞬目，口の動きなどが検討されているが，決定的な特徴は未だ定まっていない．そのため単一の特徴量からではなく，複数の特徴量を組み合わせて推定精度を向上させることが有効であると考えられる．

さらにそれらの身体特徴は内的要因で表出することもあれば，ＴＶコンテンツに含まれる特徴など，外的要因で表出する場合もある．確度の高い内部状態推定のためには，内的要因のみならず外的要因も考慮する必要がある．

4.2.3　TV視聴行動調査

本研究に先立ち，NHK放送技術研究所で行われた視聴者の家庭におけるテレビの視聴行動に関する調査結果を紹介する．本調査実験は家庭にカメラを設置し，一般家庭で実際にどのようにＴＶが見られているかを調査したものである．

本調査は以下の内容で実施した．

【方法】　：　調査者によるセルフレコーディング形式
【記録日時】　：　5日間（金曜～火曜）の間の任意
【記録場所】　：　調査者の自宅（テレビが置いてある部屋）
【記録時間】　：　平日・休日それぞれで，2時間以上の時間を記録
【記録方法】　：　2台のカメラによる定点動画記録
【記録内容】　：　普段の生活の中で，ルーティンとしてテレビをつけている時間帯をヒアリングし，なるべくその時間帯の記録をお願いした

本実験により，以下のような知見が得られた．まず，行動観察調査と調査実験後の回顧インタビュー調査の結果から，ＴＶの視聴スタイルとして，以下の4項目が抽出された．これらは，視聴者のＴＶとの関わり方や番組に対する執着性や嗜好性，家庭環境や生活パターンなど，ＴＶ視聴が生活の中でどのように位置づけられているのかを理解する上で重要な指標となり得る．

①視聴関与・・・計画視聴と非計画視聴（未計画視聴と無計画視聴）
②視聴態度・・・集中視聴とながら視聴（なにげ視聴とながし視聴）＋ 非視聴
③視聴人数・・・単身視聴と複数視聴
④視聴タイム・・・リアルタイム視聴とタイムシフト視聴

上記を図4.4にまとめる．

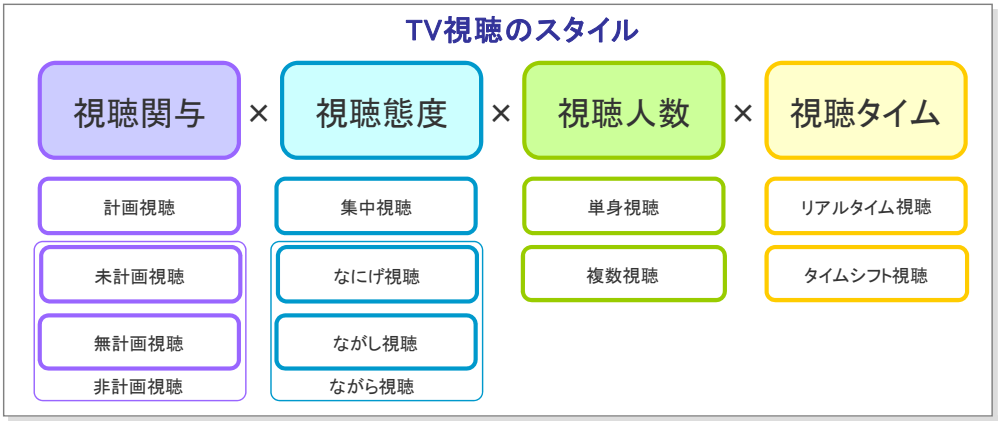


図4.4　ＴＶ視聴のスタイル

続いて，“視聴態度”の中の「集中視聴」と「ながら視聴」に着目し，それぞれについてまとめる．

集中視聴は，番組の内容に対するコミットが高く，同時行動をなるべく発生させず，視聴に意識を集中させてテレビを観ている状態を指す．一方，ながら視聴は，なにげ視聴とながし視聴に2つに区分される．なにげ視聴は，同時行動を伴うものの，番組に対しても意識を振り分け，番組内容の把握のレベルもある程度高く，興味を惹く内容（ヒットポイント）を待っている視聴状態を指す．また，ながし視聴は，同時行動に関する意識が大部分を占め，番組内容の把握は希薄で，ヒットポイントに対する反応も弱い．これらを表4.1にまとめる．

表4.1　集中状態と非集中状態

	集中状態	非集中状態
目線・顔状態	目線が**TVの方向**を向いている。 ある一定時間、（顔、目線が）**静止状態**になる	目線がTVの方だけに固定されず、**いろいろな方向**を向く。 **運動状態**が続く
身体・位置状態	非集中状態より、顔、体が前のめり、もしくは**TVの近距離**になる傾向	集中状態より、顔、体が後背、もしくは**TVから遠距離**になる傾向
連動タイミング・連動	コンテンツ内容と連動し、「一定時間の静止→リアクション」を**複数が同じようなタイミング**で繰り返す。	それぞれのうごきに**連動性がなく**、コンテンツ内容との連動性もない

＜集中視聴＞

- 番組の内容に対して強い興味・関心がある，主体的な視聴態度
- 内容が理解できなくなること，観たい瞬間を見逃すことによる満足度の低下を避けている
- 番組の内容が期待に応えるものであることを強く期待している
- 選択した番組の時間帯を視聴のみで消費する
- 中心視野型の視聴．トータルな視聴
- テレビの前で静止して食入る様に視聴する（ただし，ぼんやりしながら，うとうとしながらの場合，集中視聴ではないが，静止する）
- 平日の夕方～深夜にかけて，休日などの自由行動が取れる時間帯
- ドラマ，ドキュメンタリーなど，ストーリー性のある内容．観るポイントが途切れないスポーツ（サッカーなど）

＜ながら（なにげ）視聴＞

- 番組の内容に対して弱い興味・関心とある程度の期待がある，受動的な視聴態度
- 番組の内容を浅くサーチすることによって満足度の上昇に期待している
- 新たな興味，関心が想起される可能性をゆるく期待している
- 選択した番組の時間帯をテレビ視聴と同時行動で消費する
- 周辺視野型の視聴．聴覚などが視覚より優位になることがあるパーシャルな視聴
- テレビの前で，生活必需行動，社会生活行動，自由行動をしながら視聴し，強く興味を惹いた内容（ヒットポイント）だけをピックアップする．場合によっては，集中視聴に移行する
- 朝や夕方など，生活必需行動や社会生活行動が伴う時間帯
- バラエティ，報道・ニュースなど，単発性，速報性，一過性の内容．落ち着いて観られるスポーツ（野球，相撲など）．また，録画メディアに記録された番組もなにげ視聴される場合もある

この調査により，視聴者がＴＶコンテンツを集中して観ている時は

「目線がＴＶの方を向き，一定時間静止状態になる」

という知見が得られた．身体動作の静止については言及されていないものの，頭部が静止状態になれば，全身もそれにつれて静止状態になると予想される．

また視聴者が集中状態にある際は，身体動作がコンテンツと連動するという知見も得られた．これらのことから，視聴者の集中度（興味度）推定には身体動作の計測が有効であり，また身体特徴とコンテンツ特徴の連動性を考慮すべきであることを確認した．

4.3　提案手法

4.3.1　仮説の設定

○　注目状態の定義

視聴者がＴＶコンテンツを視聴中に抱く感情や興味はユーザの内部状態であり，外からの観測は難しい．しかし情動として身体に表出することも多く，身体動作を伴う筋運動系情動計測により，内部状態を推定できる可能性がある．

一般に表情は比較的大きな特徴変化であり，表情から感情を推定することもできる．しかし表情はユーザによるコントロールが容易な情動であり，必ずしも表情による特徴変化で正確な内部状態推定ができるとは言えない．例えばＴＶコンテンツに対して「怒り」という感情を抱いたとしても，その表情変化を自己抑制する場合もあれば，冷笑など別の感情に近い表情を作る場合もある．

そこでユーザのコントロールが比較的困難な，微小な身体特徴変化に焦点を当てることとした．微小な身体変化は，表情などに比べて映像からの検出が困難ではある．しかしそれらの動作はユーザが無意識のうちに行っている可能性が高く，より真の内部状態を表していると考える．

本研究では，“注目状態”を「身体に表出する集中状態であり，視覚情報で判断可能であるもの」と定義した．集中状態は人物の内面であり，観測不能であるのに対し，注目状態はユーザの微小動作を観測することで，客観的な観測が可能であるという位置付けである．

○　注目状態における仮説

視聴者が注目状態にある時と非注目状態にある時の情動を検討し，上記 Kinect センサと視聴者カメラ映像で観測可能な情動に関する仮説を立てた．

まず関連研究の結果の通り，視聴者が注目状態に入るとそれまでの行動を停止し，体を静止させると予想した．そこで「注目状態では身体動作が減少する」との仮説を立てた．身体動作のわずかな変動量は Kinect センサで計測可能である．この特徴量は視聴者カメラのフレーム差分量などで代用することも可能であるが，Kinect を用いることで背景ノイズを排除した，より確度の高い観測を行うことができる．

次に，先行研究より[Nakano09, Seki02, Ishiyama94, Yamamoto04]，視聴者が注目状態にある時はまばたきの回数が減ることが予想される．そこで「注目状態では瞬目数が減少（＝ 瞬目間隔が増加）する」との仮説を立てた．瞬目は視聴者カメラ映像でも確認できるため，高精細映像であれば自動検出も可能であると予想した．

また様々な先行研究で集中状態の検出に視線が用いられており[Sawahata08，Just76, Emery00]，視線は人間の内部状態を計る上での重要な指標になり得ると考えられる．注目状態時にはＴＶ画面を固視する傾向があるため，「注目状態では視線変動が減少する」との仮説を立てた．

○　外的要因（コンテンツ特徴）の影響

注目状態では各種情動が減少するという仮説を立てたものの，外的要因によってもその動作量は変化すると考えられる．たとえば，視聴中のコンテンツ映像に字幕や人物が多数存在する場合，字幕や人の表情を読むためユーザの視線移動量は必然的に多くなる．一方でユーザが呆け状態にある場合，字幕が多いシーンであってもユーザの視線は変動しない．また衝撃的なシーンを視聴した場合，驚いて思わず大きく体を揺らすことも考えられる．このように，外的要因（コンテンツ特徴）がユーザの眼球運動や身体行動にもたらす変化を考慮する必要がある．

上記 3 つの情動の中で，最もコンテンツ特徴の影響を受けると考えられるのは視線変動である．映像中の字幕量や人物の数，映像変化量などの情報が多ければ多いほど，多くの映像内オブジェクトへ視線を向ける必要が生じる．そのため，情報量が多いコンテンツを集中（注目）視聴した場合，視線変動量は増加すると予想される．これは先の仮説に反するが，コンテンツの情報量に応じて処理を分岐することで対応することとした．

以上，視聴者が注目状態にあるときは

1. 身体動作が減少する
2. 瞬目数が減少（＝瞬目時間間隔が増加）する
3. 視線変動量が減少する

3’ ただしコンテンツの情報量が多い場合は視線変動量が増加

の 4 つの仮説を立てた．

4.3.2　実験条件

上記仮説を検証するため，家庭でのＴＶ視聴環境を模した実験を行った．一般家庭でのＴＶ視聴においては，ユーザがヘッドマウント機器や生体信号測定器などの接触型デバイスを用いることは少ない．家庭での状態に近い環境で実験を行うため，接触型センサを利用せず，図 4.5 のように通常の HD カメラと Kinect センサのみでユーザの視聴状況を計測した．

ユーザの座る椅子は指定されるものの，上半身の動きには制限を設けず，腰のひねりや伸び動作など，実験を意識せず自由に動いてよいものとした．図 4.6，4.7 に実験時の視聴者カメラ映像例，Kinect での頭部・頚部位置計測例を示す．

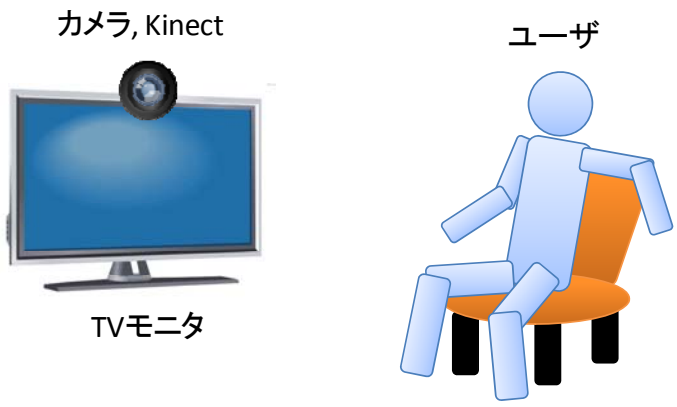


図 4.5　家庭内でのＴＶ視聴環境例

図 4.6　視聴者カメラ画像例

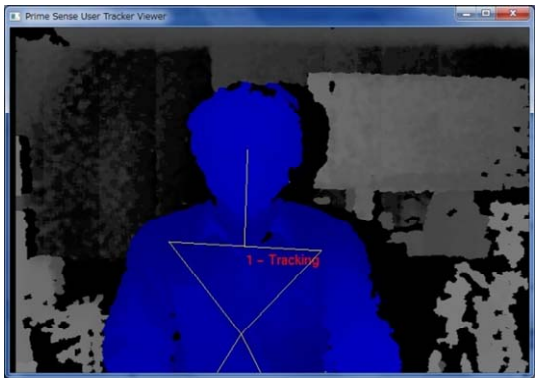


図 4.7　Kinect での頭部・頚部位置計測例

対象コンテンツはニュース番組とした．ニュース番組は番組内に政治や経済，スポーツなど様々なトピック（項目）が混在しており，注目状態と非注目状態との差が生じやすいと考えられる．トピックとは番組を話題によって意味的に分類したものである．例えばニュース番組においては，政治，経済，スポーツ，気象情報，などのトピックに分かれる．ニュース番組における各トピックの長さは 30 秒から 2 分程度であり，今回定めた特徴量から注目状態を判定するのに適度な長さであると考えられる．

トピック単位ではなく番組単位で注目度を推定した方が精度は高くなると予想されるが，コンテンツ内の短期的な興味の抽出が困難となる．一方，推定単位を数秒に設定した場合，短期的な興味の抽出は可能となるものの，ノイズが混入しやすく高い推定精度は望めない．短期的興味の抽出と推定精度の両面を考慮し，番組の内容的区切りであるトピック毎に注目度を推定することとした．

実験に使用した番組はＮＨＫニュース 7（30 分）である．コンテンツの重複視聴を避けるため，実験日当日のニュース 7 で実験を行った．ニュース 7 では 1 番組あたり 20 程度のトピックが放送される．番組視聴後，各トピックへの注目度を 1 点から 5 点（全く注目しなかった，あまり注目しなかった，どちらでもない，やや注目した，かなり注目した）で評価し，この主観評価を正解データとして用いた．被験者は 5 名とした．表 4.2 に被験者の年齢・性別と実験した番組数を示す．

表 4.2　被験者と実験番組数

被験者	A（30代男）	B（50代男）	C（30代女）	D（50代男）	E（40代女）
番組数	6	3	1	1	1
（内、視線計測器利用）	3	0	0	0	0

なお被験者 A に対しては 3 番組でキャップ型視線計測器（NAC 製 EMR-9 [NAC]）を用いた実験を行い，番組視聴時の注視点の正解データを計測した．キャップ型視線計測器装着時の例を図 4.8 に示す．キャップ中央にあるカメラで被験者正面の映像を収録し，その映像上での注視点をリアルタイムに記録できる．計測した注視点の例を図 4.9 に示す．図中の+，□マークが左目，右目の注目点を表す．たとえば，図の例では「九州」の字幕を注視している．

図 4.8　視線計測器

図 4.9　視線計測器での視線データ取得例
（モニタ内の＋，□マークが注視点）

4.3.3　仮説の検証

特徴量の取得に先立ち，正解データに基づく仮説の検証を行った．計測した 5 人分の Kinect データ，視聴者映像から各特徴の正解データを作成した．以下に正解データの取得方法を示す．

- **身体動作量**

Kinect は精度の高いセンサであり，そのまま正解データとして利用できる．そこで身体動作特徴量として，Kinect で取得した頭部位置の変動量を計測した．頭部の 3 次元位置を 2 次元画像座標上に射影し，単位時間（フレーム）間隔の変位量を算出した．この計測値を式(4-1)の$K_k(t)$で表す．

ここで k は番組内トピックの ID 番号とし，$P_x^h(t)$は時刻 t フレームにおける 2 次元座標上での水平方向の頭部位置，$P_y^h(t)$を垂直方向の頭部位置とする．またトピック k でのフレーム長を T とする．$\mathrm{K}_k(t)$をフレーム長 T で平均化し，トピック k での平均身体動作量μ_k^Kを得る．

$$K_k(t) = \sqrt{(P_x^h(t) - P_x^h(t-1))^2 + (P_y^h(t) - P_y^h(t-1))^2} \tag{4-1}$$

$$\mu_k^K = \frac{1}{T}\sum_{t=0}^{T} K_k(t) \tag{4-2}$$

- **瞬目間隔**

視聴者カメラ映像を目視カウントすることにより，被験者の瞬目タイミング $t(n)$を記録した．ここで n はトピック内の瞬目の ID 番号とする．隣接する瞬目時刻の差をとり，瞬目間隔$\mathrm{B}_k(n)$とした．後にトピック k 内の瞬目数 N で$B_k(n)$の平均をとり，平均瞬目間隔μ_k^Bを算出した．

$$B_k(n) = t(n) - t(n-1) \tag{4-3}$$

$$\mu_k^B = \frac{1}{N}\sum_{n=0}^{N} B_k(n) \tag{4-4}$$

- **視線変動量**

視線変動量の正解データには視線計測器による精度の高い注視点データを活用した．視線計測器で実験を行ったのは被験者 A だけであるが，3 番組分と多くの計測を行ったため，一定の信頼性を有すると考える．

視線計測器からは，フレーム毎に視野カメラ画像内の注視点の 2 次元座標が計測される．この座標のフレーム間の変動量を計測し，視線変動量とした．具体的には式(4-5)に沿ってトピック k のフレーム t-1, t 間における視線変動量$E_k(t)$を算出する．ここで$P_x(t)$は時刻 t における x 座標上の注視点とする．

その後，フレーム長 T で平均化し，トピック k の平均視線変動量 μ_k^K を求める．

$$E_k(t) = \sqrt{(P_x(t) - P_x(t-1))^2 + (P_y(t) - P_y(t-1))^2} \tag{4-5}$$

$$\mu_k^E = \frac{1}{T}\sum_{t=0}^{T} E_k(t) \tag{4-6}$$

上記により，トピック毎に各特徴量の平均値 μ_k^K，μ_k^B，μ_k^E が得られる．各トピックには被験者の主観により，1～5 点の注目度が与えられている．この主観評価点と各特徴量の平均値の相関値を算出し，仮説を検証した．表 4.3 にその結果を示す．絶対値が 0.2 を超えた項目に関しては赤字で示した．

身体動作に関しては，いずれの被験者においても負の相関値となった．特に被験者 A, B, C においては-0.25 を下回っており，身体動作と主観評価の間には弱いながらも逆相関の関係があると考えられる．すなわち，「注目度が高いほど身体動作が減少する」との仮説に合致した結果となった．

瞬目間隔に関しては，いずれの被験者においても正の相関値となった．特に被験者 B の相関値は 0.5 を超えており，高い正の相関が見られる．この結果は，「注目度が高いほど瞬目時間間隔が長くなる」との仮説に合致した結果となった．

表 4.3　各特徴量と主観評価値との相関

被験者（データ量）	身体動作	瞬目間隔	視線変動
A（3日）	−0.280	0.257	−0.264
B（3日）	−0.268	0.551	−
C（1日）	−0.256	0.245	−
D（1日）	−0.136	0.179	−
E（1日）	−0.042	0.271	−

視線変動に関しては，相関値-0.264 と弱い負の相関が確認された．この結果は「注目度が高いほど視線変動が減少する」との仮説に合致している．

表 4.3 より，視聴者観測で得られた各特徴量と注目度の間には，弱いながらも仮説に沿った相関がみられた．これにより，設定した仮説 1～3 の有効性が示された．

4.3.4　コンテンツ特徴の検討

4.3.1 で前述した通り，コンテンツに含まれる情報量が多い場合は仮説 3 に反して視線変動量が増加すると予想される．そこで，視線誘導性が高いコンテンツ特徴と実際の視線変動との関係を検証し，抽出すべきコンテンツ特徴を検討した．

検討にあたっては視線計測器で視点を計測した被験者 A のデータを用いた．また呆け状態の時のデータを省くため，主観評価値が 4 または 5 のトピックのみを使用した．視線の誘導性が高いと考えられるコンテンツ特徴としては，字幕数，人物数，映像の動き量の 3 つを検討した．これらコンテンツ特徴と視線変動量に相関がみられれば，「情報量が多いコンテンツでは視線変動量が増加する」という仮説 3´を証明できる．

- **字幕数と視線変動の関係**

まず初めに，コンテンツ内の字幕量と視線移動量の関係を調査した．字幕量の計測にあたっては，画面に表示された字幕数を目視カウントした．1 秒あたりの字幕数と視線計測器で計測した視線変動量の関係を図 4.10 の散布図に示す．図中の各点が 1 つのトピックに相当する．

図より，トピック内の字幕が増えるに従い，視線変動量が増加していることが分かる．相関値も 0.48 と比較的高い値が得られた．この結果から，字幕には強い視線誘導性があることが分かった．

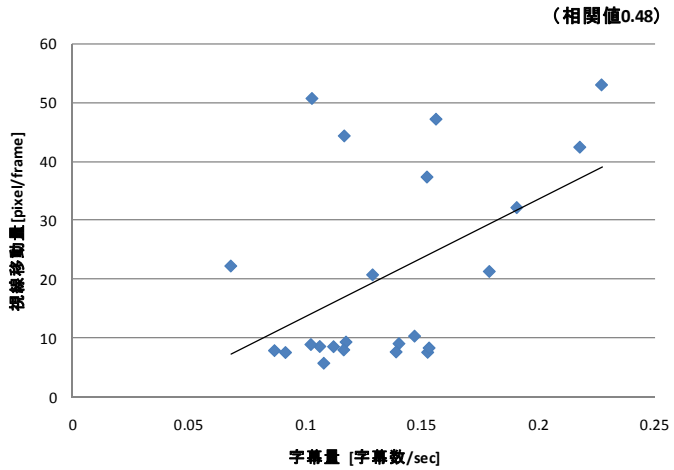


図 4.10　字幕量と視線変動の関係

・　人物数（顔の数）と視線変動の関係

次に，コンテンツ内の人物数と視線変動量との関係を調査した．映像内の人物数は，Viola & Jones の手法で画面内の顔を検出することによって自動的にカウントした[Viola01]. この顔検出アルゴリズムは OpenCV ライブラリ[OpenCV00]などを用いても実装できる．検出した顔の総数をトピック時間長で割ったものを顔の量と定義し，視線変動量との関係を分析した．

顔の量，視線変動量を軸とした散布図を図 4.11 に示す．両者の相関値は 0.37 であった．この結果より，字幕量ほどではないものの，人物数も視線変動に影響を及ぼす視線誘導性があることを確認できた．

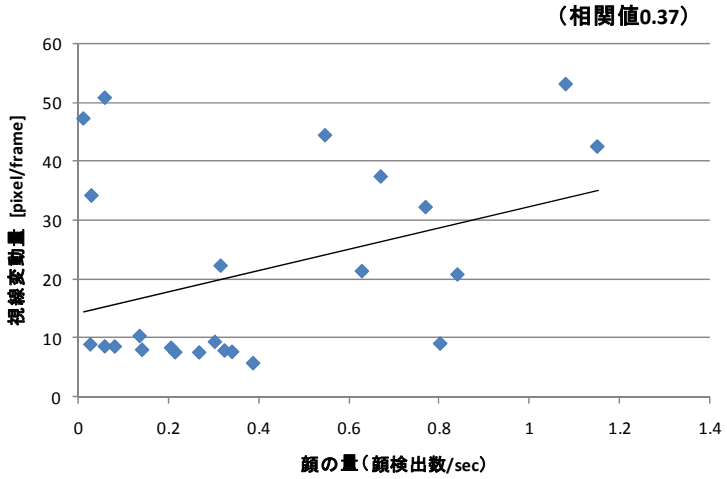


図 4.11　顔の量と視線変動の関係

・　映像の動き量と視線変動の関係

動物は自らの意思に関わらず，動いている物体を目で追う習性がある．そのため動きの多い映像では，注目状態であっても視線変動が増えることが予想される．そこで最後に，映像の動き量と視線変動量の関係を調査した．

映像の動き量は，隣接フレームでの輝度差分の平均値とした．映像の動き量とそのトピックでの視線変動量の関係を図 4.12 に示す．相関値は-0.15 であり，両者の間に相関は見られなかった．この結果から，映像の動き量は視線変動量にそれほど影響を与えないことを確認した．

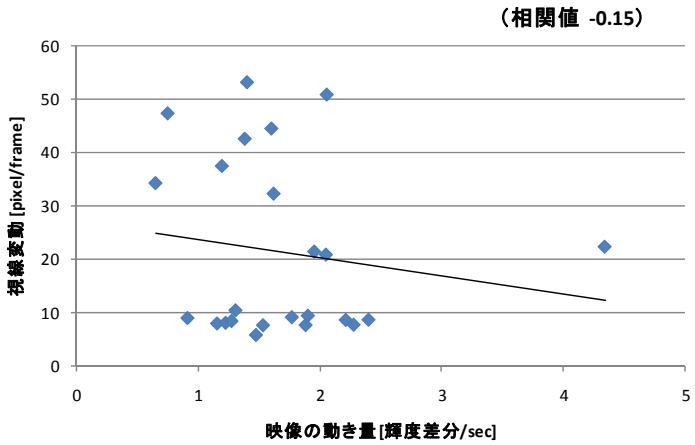


図 4.12　映像の動き量と視線変動の関係

上記 3 種類のコンテンツ特徴の検討により，字幕量や人物数が多いトピックでは必然的に視線変動量が増えることが確認された．これら 2 つのコンテンツ特徴を総じて“情報量”と呼ぶこととする．

表 4.4 に被験者 A が視線計測器を用いて実験したデータにおいて，情報過多のトピックを除外した場合の視線変動と主観評価の相関値を示す．表より，情報過多のトピックを除外することで推定精度が向上することが分かる．この結果を踏まえ，情報過多のトピックでは視線変動特徴を除外して注目度推定を行うこととした．なお，情報過多のトピックの判別法は 4.3.6 で述べる．

表 4.4　情報量過多のトピックを除外した場合の視線変動と主観評価の相関値

	相関値
全トピック	-0.264
情報過多 トピックを除外	-0.326

4.3.5　注目度推定フロー

仮説に基づき，被験者，コンテンツから特徴量を計測して注目度を自動推定する手法を考案した．図 4.13 に手法全体のブロック図を示す．本手法は，視聴中の映像コンテンツに対する視聴者の注目度をトピック毎に推定する．また注目度は離散値ではなく，連続値で出力することが可能である．

本手法への入力は Kinect センサデータ，視聴者カメラ映像，コンテンツ映像である．

いずれも家庭内汎用機器で取得可能な情報であり，研究用途の専用デバイスを用いる必要はない．本手法は図 4.13 のように特徴量抽出，特徴量記述，注目度推定の 3 ステップからなる．

まず特徴量抽出ステップでは，Kinect センサデータから身体動作量，視聴者カメラ映像から瞬目間隔と視線変動量を計測する．瞬目間隔と視線変動を精度よく計測するためには視聴者カメラ内の顔領域が正体している必要がある．そこで Kinect データを利用して視聴者の顔領域の傾きを補正した．またコンテンツ映像を解析し，映像内の情報量を算出する．この情報量に応じて特徴量記述ステップ以降の処理を分岐し，注目度推定の高精度化を図った．具体的には番組全体での情報量を基準とし，各トピック内で検出した情報量が大きい場合には，身体動作と瞬目間隔のみから注目度を推定した．

続いて特徴量記述ステップでは，特徴量抽出ステップで取得した各特徴量から特徴量記述子を作成する．特徴量の平均，分散に基づくグローバル特徴と，特徴量の頻度ヒストグラムを特徴量記述子とした．頻度ヒストグラム作成の際は，番組全体での特徴量の平均，標準偏差を基にヒストグラムのしきい値を定めた．またこの際，トピックを全体，1/2，1/4 と分割し，それぞれで特徴量記述子を作成した．さらに前後のトピックにおける後部 1/4，前部 1/4 の特徴量記述子も利用し，頑健な注目度推定を目指した．

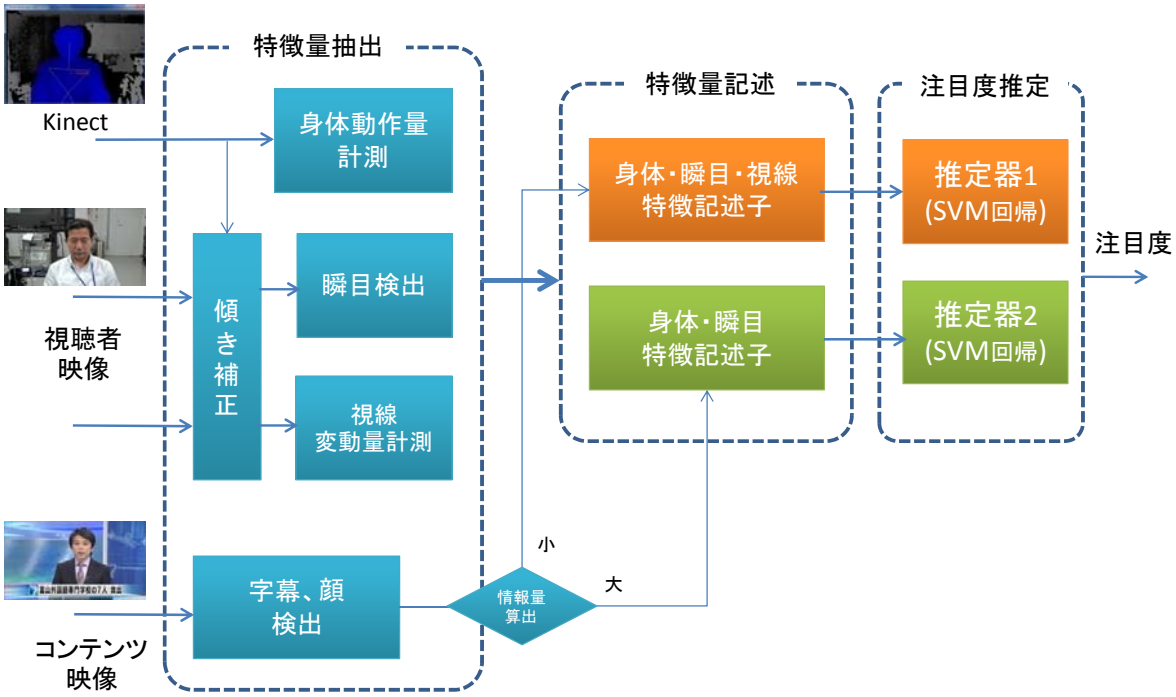


図 4.13　注目度推定のフロー

最後に注目度推定ステップでは，特徴量記述子から注目度を 0.0~1.0 の連続値で算出する．推定には教師付き機械学習アルゴリズムを用いた．事前の学習フェーズで大量の特徴量記述子を作成し，注目度推定器を学習した．情報量に応じて特徴量記述子の次元数が変わるため，身体動作，瞬目間隔，視線変動を特徴に用いた推定器と身体動作，瞬目間隔を特徴に用いた推定器の 2 通りを用意した．運用フェーズでは新たに計測した特徴量から特徴量記述子を作成し，字幕量に応じて処理を切り替えながらユーザの注目度を推定した．

4.3.6　特徴抽出ステップ

○　身体動作量計測

身体動作特徴量は 4.3.3 同様，Kinect を用いて計測する．傾き補正処理に備え，頭部に加えて頸部位置も計測した．各部の 3 次元位置を 2 次元画像座標上に射影し，単位時間（フレーム）ごとの変位量を計測した．この計測値を式(4-7)の$K_k(t)$で表す．ここで k はトピックの ID 番号とし，$P_x^h(t)$は時刻 t フレームにおける 2 次元座標上での水平方向の頭部位置，$P_y^h(t)$を垂直方向の頭部位置とする．

$$K_k(t) = \sqrt{(P_x^h(t) - P_x^h(t-1))^2 + (P_y^h(t) - P_y^h(t-1))^2} \tag{4-7}$$

○　傾き補正

続く瞬目検出と視線変動量計測は，視聴者がＴＶモニタに対して正体した状態での映像視聴を想定している．そのため，視聴者が首を傾けた状態で映像を視聴した場合には，正確な値を計測することができない．そこで Kinect で計測した首の傾きに応じ，視聴者カメラの画像全体を回転することとした．

Kinect で計測した頭部と頸部の 2 次元画像座標上での位置を利用し，首の傾きθ_tを式(4-8)で算出する．

$$\theta_t = \frac{P_y^h(t) - P_y^n(t)}{P_x^h(t) - P_x^n(t)} \tag{4-8}$$

θ_tを利用し，画像中心(cx_t, cy_t)を中心とした回転処理を施す．式(4-9)に従い，任意の点(x_t, y_t)は(x_t', y_t')へ補正される．傾き補正の例を図 4.14 に示す．

$$x_t' = (x_t - cx_t)\cos(-\theta_t) - (y_t - cy_t)\sin(-\theta_t) + cx_t$$
$$y_t' = (x_t - cx_t)\sin(-\theta_t) + (y_t - cy_t)\cos(-\theta_t) + cy_t \quad (4\text{-}9)$$

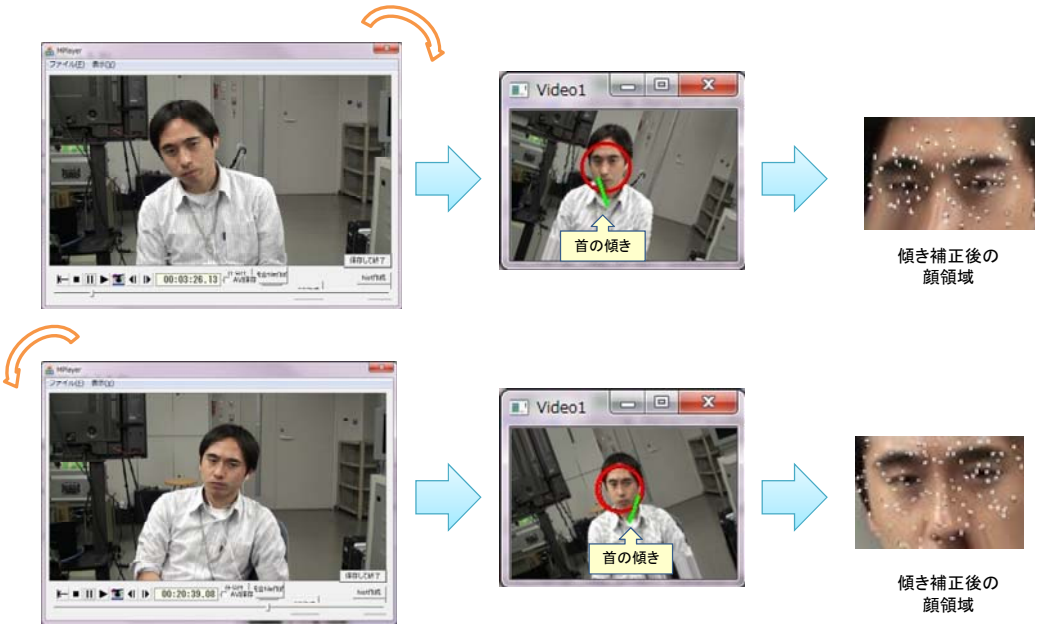


図 4.14　首の傾き補正

○　瞬目検出

前章で用いた特徴点軌跡ヒストグラムによる動作認識を応用し，瞬目検出器を作成した．はじめに，Viola & Jones の手法を用いて視聴者カメラ映像からユーザの顔領域を検出する．その後，検出した顔領域の内部で特徴点抽出・追跡処理を施し，多数の特徴点軌跡を抽出する．前章同様，各軌跡内部の動きベクトルの方向と強度に関する軌跡ヒストグラムを作成し，これを特徴量に用いて瞬目状態を判定した．

軌跡ヒストグラムは 8 通りの方向，および 0 を含む 4 通りの強度で分離した．bin 数は 25(8×3＋1)個となる．図 4.15 に軌跡ヒストグラム作成における概念図を示す．強度に関する分離閾値は 0, (0, $\bar{m}$-$\bar{\sigma}$/2], ($\bar{m}$-$\bar{\sigma}$/2, $\bar{m}$+$\bar{\sigma}$/2], ($\bar{m}$-$\bar{\sigma}$/2, ∞)と定めた．ここで$\bar{m}$は番組全体での方向ベクトル強度の平均値，$\bar{\sigma}$はその標準偏差，N は番組中の軌跡内の方向ベクトルの総数とする．

$$\bar{m} = \frac{1}{N}\sum_{i=0}^{N} m_i \quad (4\text{-}10)$$

$$\bar{\sigma} = \sqrt{\frac{1}{N}\sum_{i=0}^{N}(m_i - \bar{m})^2} \quad (4\text{-}11)$$

瞬目の判定は Bag-of-features 法（BoF 法）に基づいて行った．BoF 法は，コードブックに基づいて多次元特徴を k 種類のコードワードで代表させ，コードワードの頻度ヒストグラムで識別処理を行うアプローチである．

本件では事前に多数の軌跡ヒストグラムを作成し，この軌跡ヒストグラムを k-means アルゴリズムで k 個のクラスへ分類した．各クラスの中心をコードワードとし，コードブックへ登録した．本手法では，k の値は 100 とした．

このコードブックで量子化した bin 数 k のコードワードヒストグラムを特徴とし，2 値 SVM 識別器を学習した．図 4.15 に学習処理の流れを示す．学習データとして被験者 A の視聴者カメラ映像から瞬目時の特徴点軌跡 1,287 個，および非瞬目時の特徴点軌跡 1,714 個を用いた．この SVM 識別器により，瞬目を自動検出した．

瞬目検出器で検出した瞬目の時間間隔を計測し，注目度推定のための特徴量$\mathrm{B}_k(n)$とした．ここで k はトピックの ID 番号とし，n はトピック内の瞬目の ID 番号とする．

$$\mathrm{B}_k(n) = t(n) - t(n-1) \tag{4-12}$$

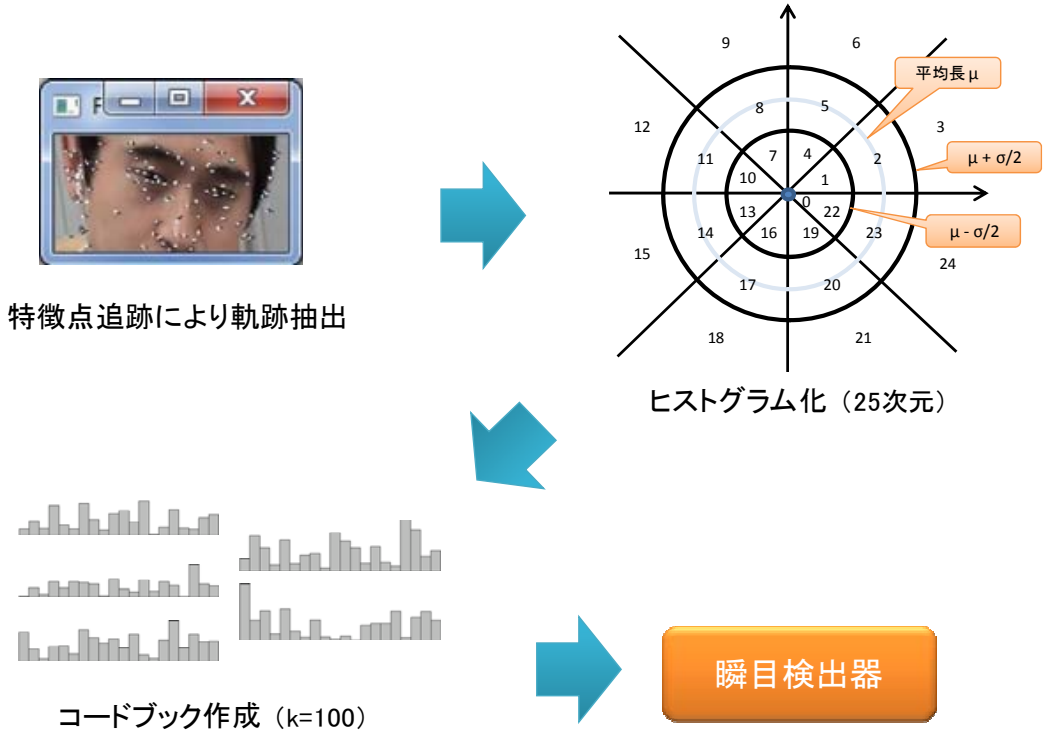


図 4.15　顔領域の特徴点軌跡ヒストグラム化

瞬目検出器の性能を確認するため，5 人の被験者データで検証実験を行った．自動計測した瞬目タイミングと目視による正解データを比較することで，瞬目検出器を評価した．各被験者カメラの映像 30 分間から瞬目を目視カウントし，目視の瞬目タイミングを正解データに用いた．ただし，瞬目検出器では検出判定が数フレームにわたり連続する傾向がある．そのため，評価にあたっては 15 フレームのマージンを設けた．この結果を表 4.5 に示す．

全被験者で総じて高い再現率が得られた．また F 値としても全員が 60%を超え，平均値は 68.2%となった．特に被験者 A では 80%を超える F 値が得られた．この結果から，瞬目検出器の有効性が示された．今回は 1 人のデータのみで識別器を学習したが，学習人数を増やすことで，より精度が向上すると考えられる．

表 4.5　瞬目検出器の精度評価

被験者	適合率 (%)	再現率 (%)	F値 (%)
A	72.93	93.72	82.03
B	51.27	80.44	62.63
C	45.65	95.15	61.70
D	51.62	97.00	67.38
E	56.22	83.70	67.26
平均	55.54	90.00	68.20

○　視線変動量計測

本研究ではユーザの瞳領域，特に黒目の動きに着目し，視線変動量を計測することとした．視線変動計測に関しては従来から様々な研究がなされているが，その多くは特殊な視線計測装置を利用して変動量を計測している．本件では家庭内での利用を想定し，専用機器を用いた精度の高い“視点”計測ではなく，汎用カメラ映像による“視線の変化量”計測を目指した．

強膜反射法は瞳領域に赤外線 LED を照射し，角膜（黒目）と強膜（白目）の反射率の違いを利用して被験者の視線方向を推定する手法である．精度の高い視線方向計測が行える一方，赤外線 LED を照射するため家庭内の使用には不向きである[Abe02]．

しかし近年のカメラは HD 解像度が一般的となり，ＴＶモニタ付近にいるユーザの瞳領域から黒目領域と白目領域を識別することも可能となった．そこで通常の HD カメラを利用し，強膜反射法の原理で視線方向を推定することとした．

まず，視聴者カメラ映像から Viola & Jones の手法を用いて顔領域を検出する．検出した顔領域から，さらに 2 つの瞳領域を検出する．瞳領域の検出原理は顔領域と同様，Haar-like 特徴量を用いた．Haar-like 特徴量は画像領域抽出のための有効な特徴量であり，検出したい領域に存在する多数の Haar-like 特徴量を学習することで，自由に検出対象を変更することができる．瞳領域検出の例を図 4.16 に示す．

瞳領域検出後，その中心部に水平方向に長い矩形領域を設置する．矩形領域を水平方向の中心で左右に分け，左右の領域内に存在する画素の輝度値の合計を計測する．瞳領域抽出と領域内への矩形の設置例を図 4.17 に示す．

図 4.16　瞳領域検出

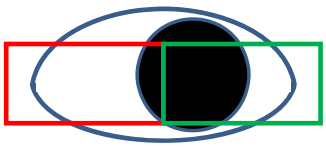


図 4.17　視線方向推定の概念図

その後，左右の領域内の輝度値の差を算出し，フレーム t での視線方向 d_t とする．この処理を式(4-13)で表す．

$$d_t = \sum_{i=0}^{N} I_l(i) - \sum_{i=0}^{N} I_r(i) \tag{4-13}$$

ユーザが左方向を向き，視聴者カメラ画像上で瞳の右側領域に黒目の割合が多くなると d_t の値は増加する．逆にユーザが右側を向き，画像上で瞳の左側領域に黒目の割合が多くなると d_t の値は減少する．汎用カメラを用いた計測であるため，視線方向推定としては十分な精度を得られないが，視線の変動量であれば，式(4-14)のように d_t の差分値

を算出することで比較的正確に計測することができる．この$\mathrm{E}_k(t)$を注目度推定の特徴量として活用する．ここでkはトピックの ID 番号を表す．

$$\mathrm{E}_k(t) = |d_t - d_{t-1}| \tag{4-14}$$

この視線変動計測器を目視観測データとの比較により評価した．被験者 A のカメラ映像 30 分間での黒目位置をマウスでトラックし，比較用正解データとした．ただし，視線変動計測器での観測は黒目の割合に関する特徴量である．両者の単位が異なるため，相関値に基づいて自動計測器の精度を評価した．

自動計測器と正解データの相関値は 0.691 であり，高い相関を確認できた．これにより，視線変動計測器の有効性が示された．

○　コンテンツ特徴（字幕・顔）検出

コンテンツ特徴として，情報量に寄与する字幕量と顔の数を視聴中の映像から自動計測する．まず字幕量を求めるため，画像中の字幕の有無を判定する識別器を作成した．字幕領域は一般に他の領域よりもコントラストが高く，エッジ特徴が表れやすい．そこでコンテンツの画像を 2 次微分であるラプラシアン画像に変換し，字幕検出のための特徴量に用いた．ラプラシアンへの変換は式(4-15)で行われる．ここで*I(x,y)*は画像*I*の(*x, y*)座標での画素値を表し，*I'*はラプラシアン画像とする．

$$I'(x,y) = 4I(x,y) - \{I(x,y-1) + I(x,y+1) + I(x-1,y) + I(x+1,y)\} \tag{4-15}$$

このラプラシアン画像は各画素で[0, 255]の範囲の値を持つ．そこで bin 数 256 のエッジヒストグラムを作成し，これを字幕検出の特徴量とした．ヒストグラムを特徴量とした識別器（2 値 SVM）を作成し，字幕の有無をフレーム毎に自動判定した．図 4.18 にラプラシアン画像の例を示す．ラプラシアン画像左下に赤色でエッジヒストグラムも表示した．この例では字幕数が多いため，右端(bin=255)での頻度が高くなっている．

字幕を検出した後，トピックkで字幕を検出したフレーム数N_kとトピックの時間長T_kの比を取り，情報量J_kを算出した．

$$J_k = \frac{N_k}{T_k} \tag{4-16}$$

最後に全トピックでのJ_kの平均$\bar{J}$と標準偏差$\bar{\sigma_J}$を算出し，$\bar{J} + \alpha\bar{\sigma_J}$ を超える情報量を持つトピックについては情報量が多いと判断した．なお，α の値は実験的に定めた．

作成した字幕検出器の性能評価を行った．学習データとは別日のニュース 7 映像から画像 180 枚を抽出し，字幕の有無を目視判定することで正解データを作成した．字幕有りと目視判定した画像は，ほぼ半数(=56%)の 101 枚であった．このデータを字幕検出器に入力し，その精度を検証した．結果を表 4.6 に示す．

適合率 0.782，再現率 0.823，F 値 0.802 と高い精度が得られた．ニュース番組における字幕はコントラストが高いため，ラプラシアンヒストグラムが有効に作用したものと考えられる．また字幕の形状がほぼ固定されていることも，高い精度に貢献したと考えられる．

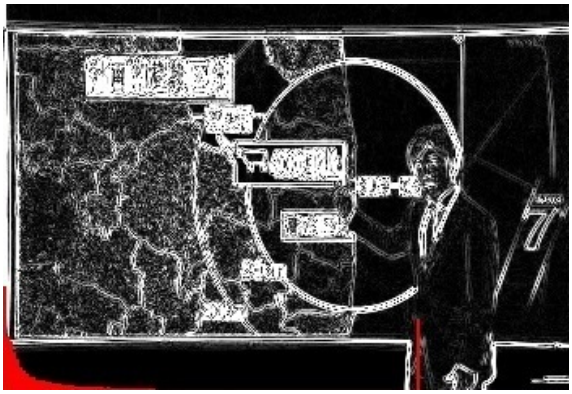


図 4.18　ラプラシアン画像の例

表 4.6　字幕検出器の性能評価

正解数	検出数	正答率	適合率	再現率	F値
101	96	0.779	0.782	0.823	0.802

字幕量に続いて，コンテンツに登場する顔の数を計測した．顔検出器には一般的な Viola & Jones の手法を用いた．フレーム毎にコンテンツ画像に顔検出処理を適用し，トピック k のフレーム分（T_k）の総数S_kを求めた．その後，トピックの時間長で正規化した値 F_k を顔の量と定めた．

$$F_k = \frac{S_k}{T_k} \tag{4-17}$$

全トピックでのF_kの平均$\bar{F}$と標準偏差 $\overline{\sigma_F}$ を算出し， $\bar{J} + \beta\overline{\sigma_F}$ を超える情報量を持つトピックについては情報量過多と判断した．なお，β の値は実験的に定めた．

4.3.7　特徴記述ステップ

上記特徴抽出ステップにて，トピック毎に 3 種の特徴 $K_k(t), B_k(n), E_k(t)$ を計測した．本特徴量記述ステップでは，各特徴を次元数固定の記述子に変換する．特徴記述子は，2 次元のグローバル特徴記述子と 8 次元のヒストグラム特徴記述子で形成される．また特徴記述子は 9 通りの時間分割パターンにおいてそれぞれ生成した．したがってトピック毎に合計 270 次元（＝3 種特徴×10 記述子×9 分割パターン）の特徴量を生成した．これを図 4.19 にまとめる．

ただし，トピック内の情報量が多いと判定された場合には，視線変動特徴量に関する記述子を除外し，180(2×10×9)次元の特徴量記述子を利用した．

特徴記述子の算出について，以下に述べる．

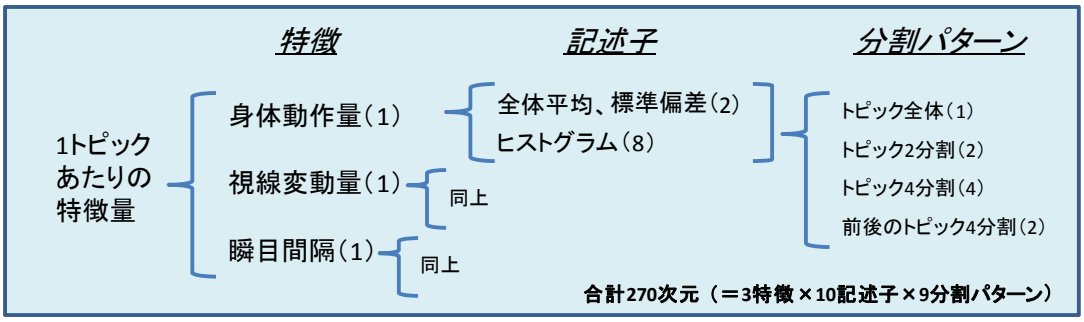


図 4.19　1 トピックあたりの特徴量

○　グローバル特徴記述子

グローバル特徴は，各特徴量のトピック内での平均値と標準偏差とした．すなわち，$\overline{K_k(t)}$，$\sigma_{K_k(t)}$，$\overline{B_k(t)}$，$\sigma_{B_k(t)}$，$\overline{E_k(t)}$，$\sigma_{E_k(t)}$ である．ここで$\overline{\mathrm{A}}$は A の平均値，σ_{A}は A の標準偏差とする．

○　ヒストグラム特徴記述子

各特徴をヒストグラム化することで，トピックの長さに依らず，特徴の生起頻度を数値化できる．頻度が特定の bin に集中しないよう，ヒストグラムのしきい値は番組全体の特徴量の平均，標準偏差により定めた．

ヒストグラムの bin 数は 8 とした．ある特徴量の番組全体での平均値を m，その標準

偏差を σ とし, そのしきい値を図 4.20 のように(-∞, *m-2σ*), [*m-2σ, m-σ*), [*m-σ, m-σ/2*), [*m-σ/2, m*), [*m, m+σ/2*), [*m+σ/2, m+σ*), [*m+σ, m+2σ*), [*m+2σ,* ∞)と定めた．各 bin のしきい値は自由に設定できるが，平均値 *m* と標準偏差 σ の値をもとに定めることで，極度に偏ったヒストグラムの生成を回避した．

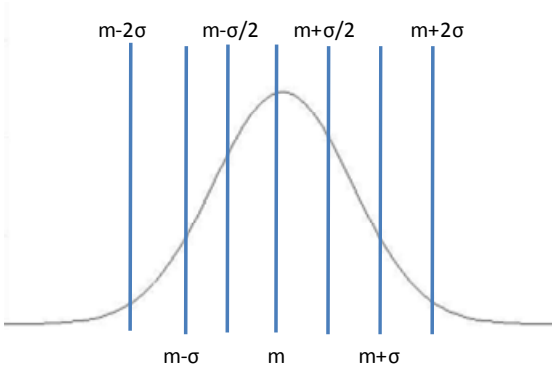


図 4.20　平均値，標準偏差値に基づくしきい値

グローバル特徴で 2 次元，局所ヒストグラム特徴で 8 次元の特徴量が得られるため，合計 10 次元の特徴記述子が得られる．

特徴記述子生成の流れを図 4.21 に示した．図では身体動作量特徴において，トピック「TPP」から特徴記述子を作成する例を示している．トピック「TPP」から特徴量の平均値，標準偏差を算出し，グローバル特徴記述子とした．また番組全体の特徴量から平均値$\bar{m}$と標準偏差$\bar{\sigma}$を算出し，それらを基にヒストグラムのしきい値を定めて 8 つのビンを持つヒストグラム特徴記述子を作成した．

○　分割パターン

ただし上記特徴量記述子はトピック全体での特徴量の生起頻度を表しているため，トピックの局所的な特徴の影響は薄れてしまう．また各種情動は，注目状態から解放された直後に増加するという知見も先行研究から得られている．そこで，トピック全体での特徴量記述子に加え，トピックを分割した場合の特徴量記述子，および前後のトピックにおける特徴量記述子も利用した．

特徴量記述子群の概念図を図 4.22 に示す．トピックを 2 分割，4 分割した場合，トピック内では 7 つの特徴量記述子が生成される．トピック前後の 4 分割ヒストグラムも加え，合計 9 種の分割パターンとした．

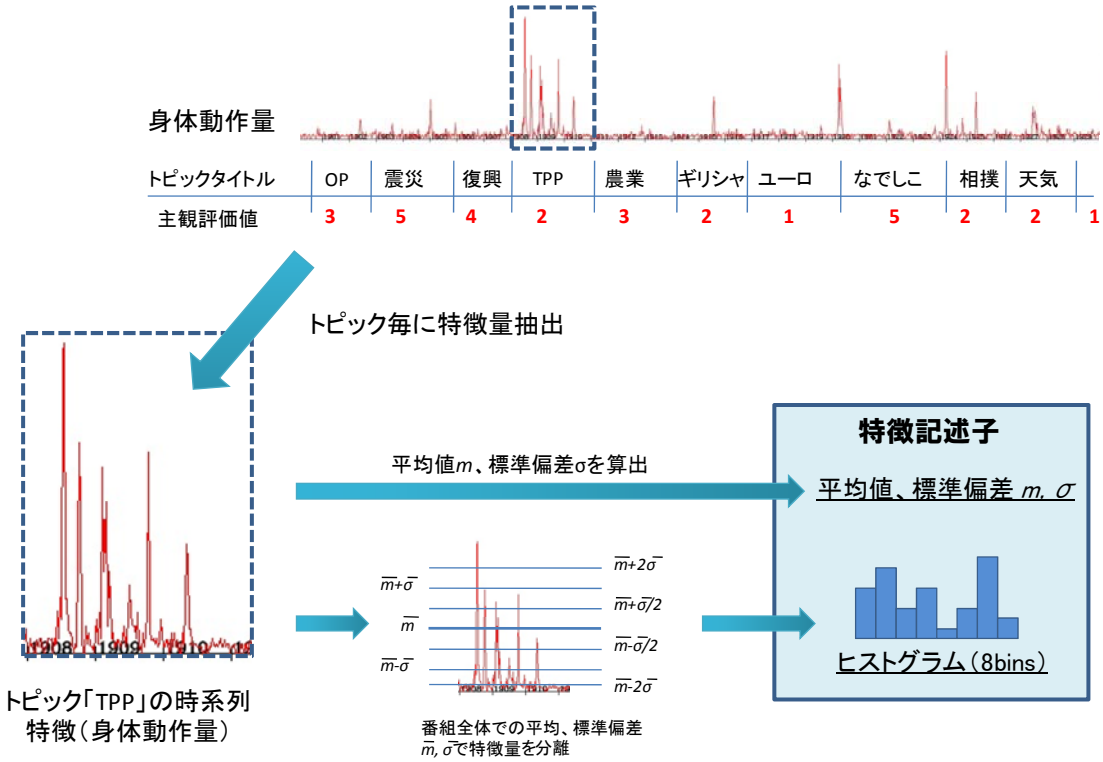


図 4.21　特徴記述子の作成

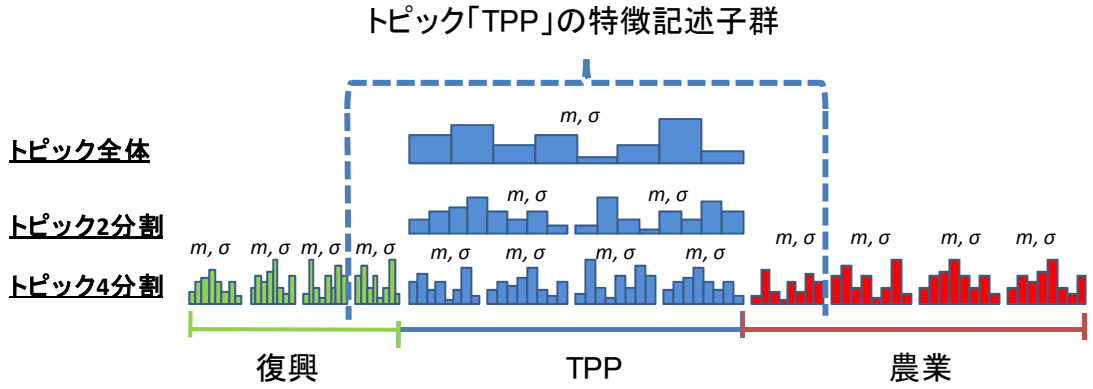


図 4.22　特徴量記述子群の概念図

4.3.8　注目度推定ステップ

注目度推定ステップでは，作成した特徴量記述子を利用してユーザの注目度を推定する．推定には SVM 回帰を用いた．SVM 回帰は，一般的な 2 値または多値の分類器とは異なり，結果を連続値として出力するアルゴリズムである．これにより，離散値ではな

いユーザの注目度合を適切に出力することができる．

事前の学習フェーズにおいて，学習用特徴量を使用して SVM 回帰推定器を作成した．図 4.23 に注目度推定器作成の概念図を示す．主観評価による正解データを付与した特徴量記述子群を特徴量とし，SVM 回帰による識別器を作成した．コンテンツ情報量による分岐により，2 通りの特徴記述子群が生成される．そこで SVM 回帰推定器も 2 つ用意した．

運用フェーズにおいては，注目度が不明なトピックの特徴量記述子群を入力とし，推定器で注目度を 0〜1 の範囲で推定した．汎用機器から得られるデータで推定を行っているため，家庭でも十分利用可能であることが本手法の特徴である．

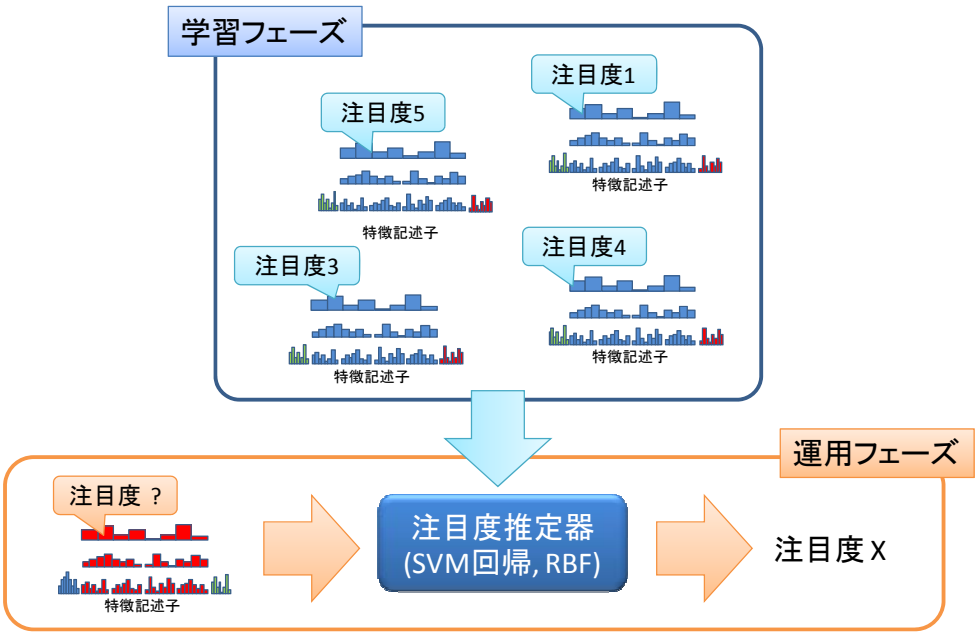


図 4.23　注目度推定器の作成

4.4　実験

4.4.1　注目度推定器の精度

作成した注目度推定器を 5 人分のデータで評価した．評価はクロスバリデーションで行い，本人以外の 4 人のデータで学習した注目度推定器を用いた評価を行った．0〜1 の範囲で出力される注目度の推定値を主観評価値と同一範囲の 1〜5 に変換し，主観評価値との平均誤差を算出した．結果を表 4.7 に示す．

各特徴量単独での評価，トピック分割なしでの評価，情報量による分岐の有無による評価など，様々な条件で推定誤差を算出した．表より，3 種類全ての特徴を利用し，か

つコンテンツの情報量による分岐を設けた場合に最も高い推定精度が得られた．この結果は 4 つの仮説の正当性を証明するものである．またトピック全体のみの特徴記述子で推定した場合は，1.483 と大きな誤差が生じた．この結果から，トピックを分割して記述子を作成することの有効性が示された．

得られた推定誤差を評価する指標として“ランダム”と“何もしない”の 2 つのベースラインを設けた．“ランダム”は 1.0～5.0 の範囲でランダムに推定した場合の主観評価との平均誤差である．“何もしない”は推定値を常に中央値の 3 と設定した場合の主観評価との平均誤差である．それぞれ 1.640，1.254 の誤差となった．

全特徴量を利用し，コンテンツを考慮した場合の推定誤差 1.305 は，ランダム推定よりも誤差が少なく，本推定器を利用することの効果が示された．しかし“何もしない”よりは低い評価結果となった．ベースラインの推定誤差は主観評価点の付け方に依存するが，今後，この値を超える精度の識別器を作成することが目標となる．

また実験に用いた視聴者映像を分析し，瞬目と視線変動に関する目視正解データを作成した．それら正解データから特徴記述子を作成し，推定器で評価したところ，表 4.7 下段に示す結果が得られた．自動計測器による推定より高い推定精度となっており，これらも各計測器の精度向上へ向けた指標となる．今後，瞬目および視線変動量計測器の精度を高め，より高精度な注目度推定器の実現を図りたい．

表 4.7　注目度の平均推定誤差

	ベースライン		1特徴			3特徴		
	ランダム	何もしない（常に注目度3）	身体動作	瞬目	視線変動	トピック分割なし	情報量分岐なし	情報量分岐あり
平均誤差	*1.640*	*1.254*	1.379	1.397	1.36	1.483	1.312	1.305

目視正解データでの推定誤差

	瞬目	視線変動
平均誤差	1.374	1.313

続いて注目の有無を 2 値判定した場合の精度を表 4.8 に示す．評価における適合率，再現率の定義は以下の通りである．

適合率：　注目度を 3 以上（または 3 以下）と推定したときに，
主観評価点が 4 か 5（または 1 か 2）

再現率：　主観評価点 4 以上（または 2 以下）のトピックを
注目度 3 以上（または 3 以下）と推定

コンテンツ情報量により処理を分岐した推定器において，F 値 0.628 の精度が得られた．この値は一般的な他の認識処理での 2 値判定精度としては高い値ではない．しかし可視情報から人間の内部状態を推定するという挑戦的タスクに対する結果としては，評価できる値ではないかと考える．この精度を 1 つのベースラインとし，今後の精度向上に努めたい．以下では，今回最も高い精度が得られた「全特徴を利用し，コンテンツ情報量により処理を分岐した場合の推定器」の結果を本推定器の結果とする．

次に，本推定器での推定値と主観評価値の相関を表 4.9 に示す．5 人の被験者それぞれの結果とデータ全体で評価した結果を示した．被験者 E については相関値 0.395 と，比較的高い相関を示した．しかし被験者 D については相関値-0.092 と，微弱ながらも逆相関の値を示した．データ全体では 0.140 の正相関となっており，注目度推定器としての有効性は示されたものの，個人差の影響は否めない．今後，個人差を考慮した特徴量取得法を検討する必要がある．

表 4.8　2 値判定での注目度推定精度

	適合率	再現率	F値
コンテンツ不使用	0.516	0.72	0.601
コンテンツ使用	0.516	0.8	0.628

表 4.9　主観評価値（正解データ）と推定値の相関

被験者	相関値
A	0.025
B	0.163
C	0.038
D	-0.092
E	0.395
データ全体	0.140

次に，主観評価点に対する注目度推定値の散布図を図 4.24 に示す．図中の縦棒の中心点が各主観評価点における推定値の平均を表し，中心点から上下の端点への長さが標準偏差を表している．あまり差が見られないものの，評価点 1 から 5 に進むにつれ，若

干ではあるが推定値が高くなっている．この結果は，本手法の推定傾向が主観評価に沿っていることを表している．

ただし主観評価 1 や 3 において，5 に近い誤推定もいくつかみられる．このような極端な誤推定を抑制する特徴量や処理フローを今後検討したい．また，本研究では注目度の正解データに視聴者本人の主観評価値を使用したが，その配点に個人差が含まれている可能性もある．次項ではこの点も含め，個人差に対する検証を行う．

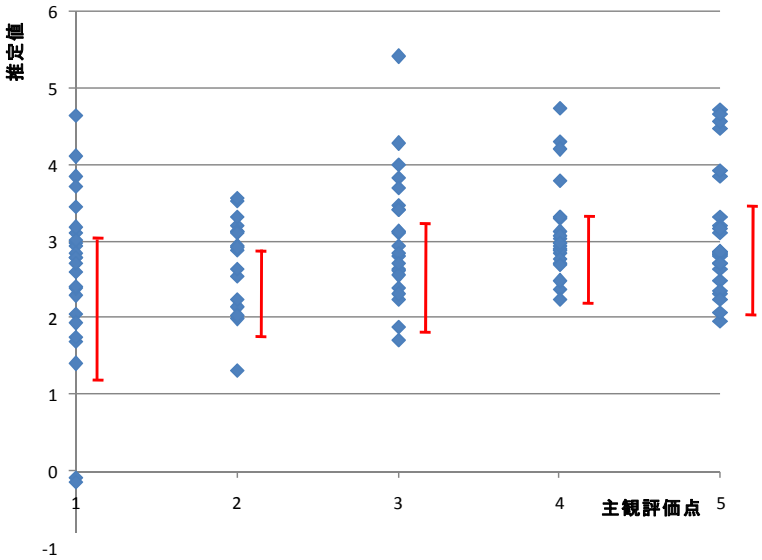


図 4.24　主観評価点と推定値の散布図

4.4.2　個人差の検証

前項で述べた通り，本推定器には個人差の影響が少なからず含まれている．表 4.10 は主観評価値に対する被験者毎の推定誤差である．被験者 E，A に対しては推定誤差 1.070，1.103 と高い性能を示した．この値は 4.4.1 で述べた“何もしない”推定よりも良い結果である．ただし被験者 C に関しては推定誤差 1.721 と“ランダム”推定より悪い結果となった．本項ではこれらの個人差に対する検証を行う．

表 4.10　各被験者の推定誤差

	A	B	C	D	E	平均
平均誤差	1.103	1.365	1.721	1.265	1.070	1.305

各被験者の主観評価点を確認したところ，中央値の 3 点が多い被験者もいれば，端点である 1，5 点が多い被験者もみられた．そこで被験者毎に評価点の平均と分散を算出し，それらに基づいて主観評価値を正規化した．表 4.11 に正規化した主観評価値に対する推定誤差と，2 値判定での F 値を示す．

表より，主観評価値を正規化することで推定精度が向上していることが分かる．推定誤差はベースラインの“何もしない”推定の誤差値に近い値となり，2 値判定での F 値は 0.7 を超えた．これにより，主観評価値を正規化することの有効性が示された．

同時にこの結果は，実験方法に検討の余地があることも示唆している．主観評価のみならず，客観評価に基づく正解データなども今後検討する必要がある．

表 4.11　正規化した主観評価値での評価

	正規化なし	正規化あり
推定誤差	1.305	1.271
2値判定 F値	0.628	0.708

続いて被験者数と推定精度の関係を検証した．被験者を 3 名から 5 名に増やしながら注目度推定器を作成し，各人数でのクロスバリデーションで推定器を評価した．各人数で作成した推定器を 2 値判定の F 値で評価した結果を図 4.25 に示す．

図より，人数を増やすほど推定精度が向上していることが分かる．これは多様な特徴量で学習することで個人差が軽減され，汎化性能が向上していることを示している．今後，新たな被験者の実験データを追加し，より汎用的な推定器を目指したい．

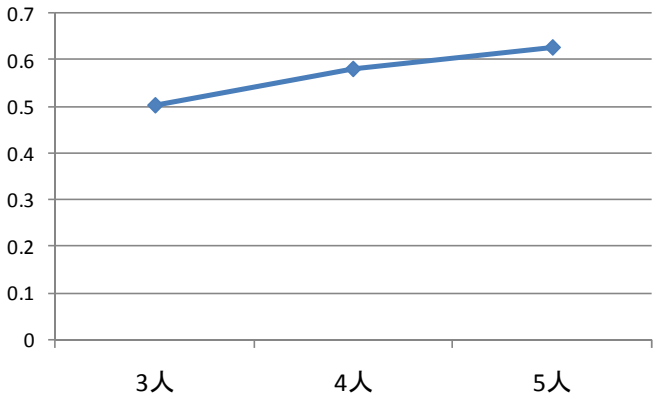


図 4.25　学習人数と推定精度の関係

4.4.3　コンテンツ特徴の利用の検討

最後に，コンテンツの情報量により処理を分岐することの効果を検証した．上記の識別器は識別器 1，2 ともに，情報量の大小に関わらず全トピックで抽出した特徴量を用いて学習した．次元数は異なるが，学習サンプルデータ数は同一である．

これに対し，情報量で分岐したトピックそれぞれでサンプルデータ集合を作成し，各データ集合のみを用いて新たに識別器 1´，2´を学習した．これらを用いて推定した場合の平均推定誤差を表 4.12 に示す．

分岐データのみで学習した場合，推定器 2´において，全データで学習した場合よりも高い精度が得られた．この結果はコンテンツの情報量による分岐が有効に働いていることを示している．また情報量が多いコンテンツでは，ユーザの情動特徴がより効果的に働いていることを示している．今後，コンテンツと情動特徴との連動性を探り，より有効な特徴量を検討することで，注目度推定器の精度向上を図りたい．

表 4.12　コンテンツ情報量の分岐データのみで学習した場合の推定誤差

	1特徴			2特徴（身体、瞬目）	3特徴		分岐データのみで学習			
	身体動作	瞬目	視線変動		分岐なし	分岐あり	推定器1'（3特徴）	推定器2'（2特徴）	推定器2'（3特徴）	推定器1',2'統合
平均誤差	1.379	1.397	1.36	1.472	1.312	1.305	1.366	1.287	1.332	1.346

4.5　まとめ

本章では，無意図的動作から人物の内部状態を推定することの可能性を検証した．具体的には，ＴＶ視聴時のユーザの微小動作（身体動作，瞬目，視線変動）からコンテンツへの注目度を自動推定した．ユーザの身体特徴に加えて映像コンテンツの特徴を考慮することで，内的要因に起因する身体特徴変化と外的要因に起因する身体特徴変化の双方を考慮した．

身体特徴の計測にあたっては，実運用を想定し，家庭環境における汎用センサを利用した．Kinect センサにより身体動作を計測し，HD カメラにより視聴者の瞬目と視線変動を計測した．目視による正解データとの比較を通し，本研究で提案する身体特徴計測器は注目度推定に有効に機能することを確認した．

各身体特徴推定器で抽出した特徴量を利用し，注目度推定器を作成した．その後，推定器の出力を被験者の主観評価値と比較し，その精度を検証した．全特徴量を用いてコンテンツの情報量により分岐した場合に最も高い精度が得られた．主観評価値との平均

推定誤差は 1.305，2 値判定における F 値は 0.628 となった．本推定器の平均誤差は注目度をランダムで推定した値を下回っており，その有効性が示された．これは可視情動の計測による，ユーザの潜在的興味度推定の可能性を示唆するものである．さらに個人差を考慮して主観評価値を正規化したところ，推定誤差は 1.271，2 値判定 F 値は 0.708 へと改善された．

今後，注目状態で表出する身体動作の個人差へ対応するため，オンライン学習などによる推定器の個人化を検討したい．また身体特徴とコンテンツ特徴とのより深い連携を考慮し，推定器のさらなる精度向上を図りたい．

第5章　結論

本論文では，実環境で撮影した映像に対する人物動作認識技術の高度化と，人物動作の観測に基づく人物の意思理解手法を提案した．家庭生活環境においても，ユーザのジェスチャによる機器操作，特定動作をクエリとした映像検索など，動作認識技術に対する期待は大きい．しかし実環境で撮影した映像は，撮影に制約が少なく，また被写体動作の多様性の問題から，高精度な解析が困難である．本研究ではこれら実環境における映像解析の課題を確認するとともに，頑健な人物動作認識技術の確立を目指した．また意図により人物動作を分類し，各段階の動作を映像解析することで，ジェスチャによるコミュニケーションからＴＶ視聴者の潜在的な興味に至るまで，それぞれの動作に含まれる意思の理解を目指した．さらに提案手法を詳細に評価し，今後の課題について検討した．

第1章では，本研究の背景として，日常生活における人物動作認識へのニーズについて述べた．人物動作の多様性を考慮し，意図の強さに基づく人物動作の分類を行った．この分類に基づき，分類した各動作への社会的ニーズを確認し，解析対象とする動作を決定した．また映像解析による一般行動認識に関する関連研究を精査し，現状の技術に対する課題を確認した．

第2章では意思伝達動作であるジェスチャに焦点をあて，その認識手法について検討した．映像解析による人物ジェスチャ認識は接触型デバイスが不要であり，次世代ＴＶ視聴環境でのインタフェースとしての期待が大きい．本章では，はじめに次世代ＴＶ視聴環境でのジェスチャ認識での要件を検討した．対話型ジェスチャの実現には奥行き情報が必須であり，また1台のセンサで奥行き計測できるデバイスが必要であることを確認した．続いてジェスチャ認識における先行研究を精査し，長期的で自然なジェスチャ認識を実現するためには，軌跡特徴の利用が有効であるとの検討を行った．そしてこれまで独立に扱われてきたセンサ依存に基づくアプローチと特徴量依存に基づくアプローチの長所を融合し，奥行きセンサと4次元軌跡特徴量に基づくジェスチャ認識手法を提案した．4次元の情報量を持つ軌跡特徴に基づくジェスチャ認識手法はこれまでになく，今後のユーザインタフェース技術に貢献するものと考える．従来手法との比較実験やシミュレーション実験を通し，その有効性を確認した．

第3章では自然な一般行動に焦点をあて，人混みで混雑した実環境での監視映像から，一般行動を頑健に認識する手法を提案した．現代では屋内・屋外問わず監視カメラが普

及しており，不特定多数の人物行動を自動認識する技術への期待が高まっている． 本章ではまず混雑映像の解析における問題点を列挙し，その課題を検討した．続いて一般行動認識の先行研究より，関連技術の現状について精査した．そして広域特徴に基づく人物追跡ベースの手法，局所特徴に基づく特徴点軌跡ベースの手法の2つを提案した．これら2つの手法の性能比較を通し，実環境で有効に機能する人物行動認識手法を検討した．混雑映像では，人物の多様な外見やオクルージョンに頑健な，局所軌跡特徴に基づく手法が有効であることを確認した．さらに，多数の軌跡をクラスタリングすることで，人物単位の行動認識を行う手法を提案し，その可能性と課題を確認した．また動作の個人差に対処するため，動作速度に不変な軌跡ヒストグラム特徴量を提案した．独自の実験に加え，映像検索に関する国際的評価型ワークショップTRECVID Surveillance Event Detectionタスクに参加し，提案手法の有効性を確認した．最後に，今後の実用化へ向けての課題を検討した．

第4章では，映像を視聴しているユーザの筋運動系情動を計測し，得られた特徴からユーザの内部状態（注目度）を推定する手法を検討した．本章では，はじめに人間の内部状態推定に関する先行研究を，脳科学や心理学の文献を含めて検討した．従来では，接触型の生体信号計測器を用いて内部状態推定がなされることが多く，非接触型デバイスでの推定は一般に困難とされていた．しかし身体動作には真の内部状態が表出するとも言われており，動作解析による内部状態推定への期待が高まっている．そこでユーザが注目状態にあるときに表出する情動を検討し，それら動作と注目度に関する仮説を立てた．また外的要因である映像コンテンツ特徴による身体動作の変化も考慮した．目視正解データにより仮説を検証し，各情動を自動計測する手法を提案した．最後に，計測した特徴から注目度を算出する注目度推定器を作成し，その性能を評価した．可視情報からの人物の内部状態推定への可能性を確認し，また今後の方向性を確認した．

本論文の成果は，実環境における人物動作認識のニーズとその実現へ向けた課題を確認し，各ニーズに対する人物動作認識手法を提案し，その有効性を検証したことである．局所特徴点軌跡ヒストグラムや，距離情報を用いた前景抽出など，実環境での課題に対処した様々な人物動作認識手法を提案した．さらに人物の意図により動作を分類し，各段階での知見を活用してより微小動作の認識手法を提案した．具体的には，意志伝達動作，一般動作，情動動作の解析にあたり，各段階で得られた知見の特徴点軌跡，軌跡ヒストグラム，特徴量ヒストグラムを活用した．本研究は，映像解析による人物意図理解へ貢献するものであり，特に，無意図的動作（情動）から人物の潜在的興味を推定する可能性を示したことは，本研究の大きな成果であると考える．本研究で提案した人物動作理解に関する技術は，家庭における新たなマンマシンインタフェースや個人プロファイル推定，監視カメラをはじめとする各種映像における人物行動検出など，様々な分野

へ応用可能な技術である．

本論文にて，実環境でも有効に機能する人物動作認識手法について述べた．しかし実用化に向けては課題も残る．中でも動作に表出する人物意図が少ない一般行動や情動に関しては，先行研究における課題は克服しているものの，実用化レベルの認識精度には未だ達していない．今後も頑健な人物動作認識の実現へ向け，抽出特徴量や識別手法を引き続き検討する必要がある．特に情動解析による人物の内部状態推定は挑戦的な課題であり，先行研究も少なく，検討すべき課題が多く残る．注目状態に表出する身体動作の個人差への対応はその一つである．また身体動作特徴とコンテンツ特徴の連携方法にもさらなる検討の余地がある．しかし身体に表出する情動からユーザの内部状態を推定する可能性を示せたことは，本論文の大きな貢献である．この研究で得られた知見を踏まえ，今後もより確度の高い人物動作認識技術を目指して研究を進めたい．

謝辞

本論文は，様々な方々のお力添えのもとに完成しました．
はじめに，本論文の主査であり，主任指導教官でもありました国立情報学研究所/総合研究大学院大学の佐藤真一教授におかれましては，本研究全般に関して多大なご指導とご鞭撻を賜りました．お忙しい中でも，いつも快く相談に乗っていただき，熱心に私の研究を支えていただきました．特に最終年度は再三にわたりご相談させていただきましたが，私の未熟な点に対しましても粘り強くご指導くださいました．この先生のご指導なくしては，本研究は成し得ないものでありました．研究以外の活動におきましても，学会等において様々な先生方に私を紹介していただくなど，研究者としての人生にも大きな糧となる経験をさせていただきました．心より厚く御礼申し上げます．

また本論文をまとめるにあたり，論文審査委員の先生方からも多大なご指導とご鞭撻を賜りました．杉本晃宏教授におかれましては，中間審査や予備審査，本審査において多くの建設的な指導をしていただき，研究をより高みへと導いてくださりました．また審査の場以外でも，私の論文を何度も精査していただき，多くの助言を授けていただきました．またメディア処理応用の授業におきましても，私に本研究の糧となる知識を与えていただきました．厚く御礼申し上げます．佐藤いまり准教授におかれましては，常に温かい目で私の研究を支えていただき，研究のみならず研究者としての心構えなど，幅広いご指導を賜りました．放送局ゆえの研究の進め方など，私の独自性を発揮できる方向性を示していただきました．またメディア処理応用の授業におきましても，専門分野外の私に対して分かりやすい言葉で説明してくださりました．深く感謝しております．孟洋助教におかれましては，研究の方向性で悩む私に気さくに声をかけていただき，何気ない言葉の中から今後の研究の道筋のヒントを与えていただきました．審査の場におきましても熱心にご指導いただき，私の研究に新たな視点を与えてくださりました．またメディア処理基礎の授業におきましても，私の拙い英語で執筆した課題に目を通していただき，ご指導いただきました．心より感謝いたします．佐藤洋一教授におかれましては，お忙しい中 NII まで足を運んでいただき，私の研究を審査していただきました．また審査の場以外におきましても，私の研究を支えていただき，論文の完成度を高めるためご指導いただきました．いただいた助言はこの論文のみならず，今後の研究生活にも大きな糧となります．心から厚く御礼申し上げます．

また古山宣洋准教授におかれましては，私の専門外である心理学の見地から私の研究の新たな方向性を示していただきました．面識のない私に対しても快く相談に乗っていただき，論文に心理学の視点を与えていただきました．厚く御礼申し上げます．

皆様のご指導，ご鞭撻により本論文をまとめ，より質の高いものとすることができました．重ねてお礼申し上げます．

また，本研究の機会を与えてくださり，ご指導，ご鞭撻を賜った日本放送協会の多くの方々に心より感謝致します．特に，放送技術研究所の久保田啓一所長所長のご理解とご鞭撻は，本論文をまとめるうえで大きな支えとなりました．厚く御礼申し上げます．八木伸行氏には，入学以前から研究内容のみならず，様々な面でご指導をいただきました．また研究者として多くの経験の場を与えていただきました．厚く御礼申し上げます．更に，柴田正啓氏，藤井真人氏，苗村昌秀氏，佐野雅規氏にも，研究へのご指導をいただくとともに，仕事と学業の両立の面で様々なサポートをいただきました．心より御礼申し上げます．サイモン クリピングデル氏，住吉英樹氏，三ッ峰秀樹氏，山内結子氏，武藤一利氏，望月貴裕氏，松井淳氏，大久保英彦氏，奥田誠氏，河合吉彦氏，古宮弘智氏，Duy-Dinh Le， 武小萌，朱才志，浜田玲子，Xiao Zhou，森靖英，Ngo Duc Thanh をはじめとした NHK 放送技術研究所の皆様，並びに国立情報学研究所/総合研究大学院大学佐藤真一研究室の皆様には，日頃のご討論，ご指導，ご鞭撻を賜りましたこと，深く感謝致します．

最後に，私を育ててくれた両親をはじめ，心の支えとなりいつも元気づけてくれた家族に感謝します．

参考文献

[Abe02] 阿部清彦, 大井尚一, 大山実, “強膜反射法を用いた画像解析による視線入力システム : 目頭抽出による頭部移動の検出と補正,” 電子情報通信学会技術研究報告. SP, 音声 102(418), 47-52, 2002.

[Ahad08] Md. Atiqur Rahman Ahad, T. Ogata, Joo Kooi Tan, Hyoungseop Kim, and Seiji Ishikawa, “View-based Human Motion Recognition in the Presence of Outliers,” Biomedical Soft Computing and Human Sciences, Vol.13, No.1, pp.71–78, 2008.

[Appenrodt10] Jörg Appenrodt, Ayoub Al-Hamadi, and Bernd Michaelis, “Data Gathering for Gesture Recognition Systems Based on Single Color-, Stereo Color- and Thermal Cameras,” International Journal of Signal Processing, Image Processing and Pattern Recognition. Vol.3, No.1, pp.37-50, 2010.

[Argyle75] Michael Argyle, “Bodily Communication,” Methuen & Col Ltd, 1975

[Bahar07] B. Bahar, I. B. Barla, Ö. Boymul, Ç. Dicle, B. Erol, M. Saraçlar, T. M. Sezgin, and M. Železný, “Mobile-phone based gesture recognition,” Proc. of the eNTERFACE’07 Workshop on Multimodal Interfaces. pp.139-146, 2007.

[Basharat09] Arslan Basharat, Alexei Gritai, and Mubarak Shah, “Learning object motion patterns for anomaly detection and improved object detection,” IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp.1-8, 2009.

[Bay08] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. "SURF: Speeded Up Robust Features," Computer Vision and Image Understanding (CVIU), Vol.110, No.3, pp.346–359, 2008.

[Beiping11] Hou Beiping, and Zhu Wen, "Fast Human Detection Using Motion Detection and Histogram of Oriented Gradients," Journal of Computers, Vol.6, No.8, pp.1597-1604, 2011.

[Bhuyan08] M. K. Bhuyan, P. K. Bora, and D. Ghosh, "Trajectory guided recognition of hand gestures having only global motions", International Journal of Computer Sciences.3, pp.222-233, 2008.

[Blank05] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri, “Actions as space-time shapes,” In Proc. of IEEE Int. Conf. on Computer Vision, Vol.2, pp.1395-1402, 2005.

[Bobick01] Aaron F. Bobick, and James W. Davis, “The recognition of human movement using

temporal templates," IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 23 (3), pp.257–267, 2001.

[Chen05] Pai-Hsuen Chen, Chih-Jen Lin, and Bernhard Schölkopf, "A tutorial on ν-support vector machines," Applied Stochastic Models in Business and Industry, Vol.21, pp.111–136, 2005.

[Chen09] Ming-yu Chen, and Alex Hauptmann, "MoSIFT: Recognizing Human Actions in Surveillance Videos," CMU-CS-09-161, Carnegie Mellon University, 2009.

[Csurka04] Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray, "Visual categorization with bags of keypoints," ECCV Workshop on Statistical Learning in Computer Vision, pp.1–22, 2004.

[Dalal05] Navneet Dalal, and Bill Triggs, "Histograms of oriented gradients for human detection," In IEEE Conf. on Computer Vision and Pattern Recognition, Vol.1, pp.886-893, 2005

[Emery00] N. J. Emery, "The eyes have it : the neuroethology, function and evelution of social gaze," Neuroscience & Biobehavioral Reviews, Vol.24, No.6, pp.581-604, 2000.

[Fathi08] Alireza Fathi, and Greg Mori, "Action recognition by learning mid-level motion features," In Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, pp.1-8, 2008.

[Fujiyoshi07] 藤吉弘亘, "Gradientベースの特徴抽出 - SIFTとHOG - ", 情報処理学会 研究報告 CVIM 160, pp. 211-224, 2007

[Ganapathi10] Varun Ganapathi, Christian Plagemann, Daphne Koller, and Sebastian Thrun, "Real time motion capture using a single time-of-flight camera," IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp.755–762, 2010.

[Gorelick07] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri, "Actions as Space-Time Shapes," IEEE Transactions on pattern analysis and machine intelligence, Vol.29, No.12, pp.2247-2253, 2007

[Grimble94] Michael J. Grimble, "Robust industrial control: Optimal Design Approach for Polynomial Systems," Prentice Hall, pp.443–456, 1994.

[Guorong96] Xuan Guorong, Chai Peiqi, and Wu Minhui, "Bhattacharyya Distance Feature Selection," In Proc. of the International Conference on Pattern Recognition, Vol.2, pp.195–199, 1996.

[Handa01] 阪田真己子, "身体表現における感性情報の認知に関する研究,"神戸大学学位論文, 甲 2499, 2001.

[Hijikata07] 土方嘉徳, "嗜好抽出と情報推薦技術," 情報処理学会誌, Vol.48, No.9, pp.957-965, 2007.

[Hilton06] Adrian Hilton, Pascal Fua, and Remi Ronfard, “Modeling people: Vision-based understanding of a person’s shape, appearance, movement, and action,” Computer Vision and Image Understanding, Vol.104, pp.87–89, 2006.

[Hirata97] 平田道憲, “生活時間研究における行動分類, ”広島大学教育学部紀要, 第二部, 第 46 号, 1997.

[Hu08] Min Hu, Saad Ali, and Mubarak Shah, “Learning motion patterns in crowded scenes using motion flow field,” In Proc. of ICPR, pp.1–5, 2008.

[Igawa10] 井川一樹, 福原知宏, 藤井秀樹, 武田英明, “テレビ番組の視聴履歴と電子番組表を用いた番組推薦システムの構築と評価,” 人工知能学会全国大会, No.3, C4-3, 2010.

[Ikemura10] Sho Ikemura, and Hironobu Fujiyoshi, “Real-Time Human Detection using Relational Depth Similarity Features,” ACCV2010, Lecture Notes in Computer Science, Volume 6495/2011, pp.25–38, 2010.

[Ishiyama94] 石山邦彦, 藤井充, 山田光穂, 村上新治, 磯野春雄, “固視中の視線の安定度と瞬目数の関係,”電子情報通信学会秋季大会予稿集, D-102, 1994.

[Just76] Marcel Adam Just, and Patricia A. Carpenter, “Eye fixations and cognitive processes,” Cognitive Psychology, Vol.480, pp.441-480, 1976.

[Kagaya05] 加賀谷拓, 羽倉淳, 藤田ハミド, “無意図的動作に着目した人間のしぐさからの情動推定手法,” 日本ソフトウェア科学会, 2005.

[Kaiho00] 海保博之, 瞬間情報処理の心理学―人が二秒間でできること, 福村出版, 2000.

[Kajiwara11] 梶原伸治, “脳波計測を用いた運転者の視覚および触覚の負担推定,” 近畿大学理工学部研究報告 47, pp.9-14, 2011.

[Kellokumpu08] Vili Kellokumpu, Guoying Zhao, and Matti Pietikäinen, “Human activity recognition using a dynamic texture based method,” in: Proceedings of the British Machine Vision Conference (BMVC’08), Leeds, United Kingdom, pp. 885–894, 2008.

[Kubota06] 久保田新, “運動・動作・行動の流れの中の「意志」と意識,” 理学療法学, Vol.33, No.4, pp.199-201, 2006.

[Kudo99] 工藤力, “しぐさと表情の心理分析, ”福村出版, 1999

[Kurita08] Teppei Kurita, and Takashi Chikayama. "Classification precision of several candidates using multi-class support vector machine in generic object recognition," Technical report of IEICE, Multimedia and virtual environment 108 (328), pp.251–258, 2008[in Japanese].

[Laptev03] Ivan Laptev, and Tony Lindeberg, “Space–time interest points,” in: Proceedings of

the International Conference on Computer Vision (ICCV'03), vol. 1, pp.432–439, 2003.

[Li08] Zhu Li, Yun Fu, Thomas S. Huang, and Shuicheng Yan, "Real-time human action recognition by luminance field trajectory analysis," In Proc of ACM Multimedia, pp.671–676, 2008.

[Lien08] Kuo-Chin Lien, and Chung-Lin Huang, "Multiview-Based Cooperative Tracking of Multiple Human Objects," EURASIP Journal on Image and Video Processing, Vol.2008, Article ID 253039, 13 pages, 2008.

[Matikainen09] Pyry Matikainen, Martial Hebert, and Rahul Sukthankar, "Trajectons: Action recognition through the motion analysis of tracked features," Workshop on Video-Oriented Object and Event Classification (ICCV), 2009.

[Matikainen10] Pyry Matikainen, Martial Hebert, and Rahul Sukthankar, "Representing Pairwise Spatial and Temporal Relations for Action Recognition," Proceedings of European Conference on Computer Vision (ECCV), 2010.

[Matsubara10] 松原孝志, 徳永竜也, 黒澤雄一, 星野剛史, 尾崎友哉, "快適操作を提供するユーザインタフェース技術", 日立評論, Vol.91, No.9, pp.48-53, 2010

[Matsumura06] 松村京子, "乳児の情動研究：非接触法による生理学的アプローチ," ベビーサイエンス, 6, 2-14, 2006.

[Messing09] Ross Messing, Chris Pal, and Henry Kautz, "Activity recognition using the velocity histories of tracked keypoints," in: Proceedings of the International Conference On Computer Vision (ICCV'09), Kyoto, Japan, September, pp.1–8, 2009.

[Mezaris10] Vasileios Mezaris, Anastasios Dimou, and Ioannis Kompatsiaris. "Local invariant feature tracks for high-level video feature extraction," Proc. 11th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2010), 2010.

[Microsoft10] Microsoft, USA. XBOX Kinect, http://www.xbox.com/kinect

[Mikolajczyk08] Krystian Mikolajczyk, and Hirofumi Uemura, "Action recognition with motion-appearance vocabulary forest," IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2008.

[Ministry of Education,Culture,Sports,Science & Technology in Japan05] "情動の科学的解明と教育等への応用に関する検討会,"報告書, 文部科学省, 2005, 10

[Mitra07] Sushmita Mitra, and Tinku Acharya, "Gesture Recognition: A Survey," IEEE Transaction on systems, man, and cybernetics-part C: Applications and reviews, Vol. 37, No.3, 2007.

[Morency06] Louis-Philippe Morency, and Trevor Darrell, "Head gesture recognition in

intelligent interfaces: The role of context in improving recognition," Proc. of the 11th International Conference on Intelligent User Interfaces (IUI), 2006.

[Morris91] Desmond Morris 著, 藤田統訳, "マンウォッチング,"小学館, 1991

[NAC] NAC Eye Mark Recorder, http://www.eyemark.jp/index.html

[Nakano09] Tamami Nakano, Yoshiharu Yamamoto, Keiichi Kitajo, Toshimitsu Takahashi, and Shigeru Kitazawa, "Synchronization of spontaneous eye blinks while viewing video stories," Royal Society B, 2009.

[Nefian01] Ara V. Nefian, Radek Grzeszczuk, and Victor Eruhimov, "A statistical upper body model for 3D static and dynamic gesture recognition from stereo sequences," Proc.ICPR2001, Vol.2, pp.286–289, 2001.

[Nickel07] Kai Nickel, and Rainer Stiefelhagen, "Visual recognition of pointing gestures for human-robot interaction," Image and Vision Computing, Vol.25, Issue 12, pp.1875-1884, 2007.

[Nintendo06] Nintendo, Japan. Wii Remote Controller, http://www.nintendo.com/wii/what/controllers#remote

[NIST] National Institute of Standards and Technology (NIST), http://www.nist.gov/index.html

[Ojala02] Timo Ojala, Matti Pietikainen, and Topi Maenpaa, "Multiresolution grayscale and rotation invariant texture classification with local binary patterns," IEEE Trans. Pattern analysis and machine intelligence, Vol.24, no.7, pp.971-987, 2002.

[OpenCV00] Open CV video library, http://opencv.willowgarage.com/wiki/

[Panasonic09] Panasonic, Japan. D-Imager, http://www2.panasonic.biz/es/densetsu/device/3DImageSensor/en/index.html

[Park08] Chang-Beom Park, Myung-Cheol Roh, and Seong-Whan Lee, "Real-Time 3D Pointing Gesture Recognition in Mobile Space," IEEE Conference on Automatic Face and Gesture Recognition, 2008.

[Perbet09] Frank Perbet , Atsuto Maki , and Björn Stenger, "Correlated Probabilistic Trajectories for Pedestrian Motion Detection,"In Proc. of ICCV,

[Plagemann10] Christian Plagemann, Varun Ganapathi, Daphne Koller and Sebastian Thrun, "Realtime identification, and localization of body parts from depth images," In IEEE Int. Conference on Robotics and Automation (ICRA), 2010.

[Poppe10] Ronald Poppe, "A survey on vision-based human action recognition," Image and

Vision Computing, 28, 6, June 2010.

[Rajesh09] V. Rajesh, and P. Rajesh Kumar, “Hand gestures recognition based on SEMG signal using wavelet and pattern recognition,” International Journal of Recent Trends in Engineering. Vol.1, No.4, pp.26-28, 2009.

[Rehg94] James M. Rehg, and Takeo Kanade, “Digiteyes: Vision-based hand tracking for human-computer interaction,” In Workshop on Motion of Non-Rigid and Articulated Bodies, pp.16-24, 1994.

[Salton88] Gerard Salton, and Christopher Buckley, “Term-weighting approaches in automatic text retrieval,” Information Processing & Management 24 (5), pp.513–523, 1988

[Salvucci01] Dario D. Salvucci, and John R. Anderson, “Automated Eye-Movement Protocol Analysis,” Human-Computer Interaction, Vol.16, pp.39–86, 2001.

[Sase10] 佐瀬巧，近藤竹雄，中川匡弘，“脳血流と脳波の局所的同時計測による感情状態判別の試み,” 電子情報通信学会技術研究報告, Vol.110, No.294, pp.57-62, 2010.

[Sawahata08] 澤畠康仁, 小峯一晃, 比留間伸行, 伊藤崇之, 渡辺誓司, 鈴木祐司, 原由美子, 一色伸夫, "番組視聴時の視線分布と番組内容理解度の関係, " 映像情報メディア学会誌 Vol.62, No.4, pp.587-594, 2008.

[Schölkopf01] Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alex J. Smola, and Robert C. Williamson, “Estimating the support of a high-dimensional distribution,” Neural Computation 13, pp.1443–1471, 2001.

[Schuldt04] Christian Schuldt, Ivan Laptev, and Barbara Caputo, “Recognizing human actions: a local SVM approach,” In Proc. of IEEE Int. Conf. on Pattern Recognition, Vol.3, pp.32–36, 2004.

[Scovanner07] Paul Scovanner, Saad Ali, and Mubarak Shah, “A 3-dimensional SIFT descriptor and its application to action recognition,” in: Proceedings of the International Conference on Multimedia (MultiMedia’07), pp.357–360, 2007.

[Seifried09] Thomas Seifried, Michael Haller, Stacey D. Scott, Florian Perteneder, Christian Rendl, Daisuke Sakamoto, and Masahiko Inami, "CRISTAL: a collaborative home media and device controller based on a multi-touch display," ITS, pp.33-40, 2009.

[Seki02] 関輝夫，”「しぐさ」でわかる相手の心理，”新星出版社, 2002.

[Shi94] Jianbo Shi, and Carlo Tomasi “Good features to track,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp.593–600, 1994.

[Shiraki06] Takayoshi Shiraki, Hideo Saito, Yoshikazu Kamoshida, Katsuhiko Ishiguro, Ryo Fukano, Tatsuya Shirai, Kenjiro Taura, Mihoko Otake, Tomomasa Sato, and Nobuyuki Otsu, “Real-time motion recognition using CHLAC features and cluster,” Proc. of IFIP

International Conference on Network and Parallel Computing (NPC), pp.50–56, 2006.

[Shotton11] Jamie Shotton, Andrew Fitzgabbon, Mat Cook, Toby Sharp, Mark Finnochio, Richard Moore, Alex Kipman, and Andrew Blake, “Real-time human pose recognition in parts from a single depth image,” In CVPR, 2011

[Sillito08] Rowland R. Sillito, and Robert B. Fisher, “Semi-supervised learning for anomalous trajectory detection,” Proc. BMVC, pp.1035–1044, 2008.

[Smeaton06] Alan F. Smeaton, Paul Over, and Wessel Kraaij, “Evaluation campaigns and TRECVid,” In Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, pp. 321-330, 2006.

[Sugawara08] M. Sugawara, “Super Hi-Vision - research on a future ultra-HDTV system,” EBU Technical Review, Q2, 2008.

[Sugimura10] 杉村大輔, 木谷クリス真実, 岡部孝弘, 佐藤洋一, 杉本晃宏, “歩容特徴と局所的な見えを用いた特徴点軌跡クラスタリングによる混雑環境下人物追跡,”電子情報通信学会論文誌, Vol.J93-D, No.8, pp.1512–1522, 2010

[Sun09] Ju Sun, Xiao Wu, Shuicheng Yan, Loong-Fah Cheong, Tat-Seng Chua, and Jintao Li, “Hierarchical spatio-temporal context modeling for action recognition,” in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR’09), pp.1–8, 2009.

[Sun09] Xinghua Sun, Mingyu Chen, and Alexander Hauptmann, “Action recognition via local descriptors and holistic features,” IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) for Human Communicative Behaviour Analysis, pp.58-65, 2009.

[Thurau08] Christian Thurau, and Václav Hlavác̆, “Pose primitive based human action recognition in videos or still images,” in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR’08), Anchorage, AK, pp.1–6, 2008.

[Tsai06] Yao-Te Tsai, Huang-Chia Shih, and Chung-Lin Huang, “Multiple Human Objects Tracking in Crowded Scenes,” International Conference on Pattern Recognition (ICPR), pp. 51–54, 2006.

[Umemoto11] 梅本和俊，山本岳洋，中村聡史，田中克己, “ユーザの視線を利用した検索意図推定とそれに基づく情報探索支援,” 日本データベース学会論文誌 Vol.10, No.1, pp.61-66, 2011.

[Valstar04] Michel Valstar, Maja Pantic, and Ioannis Patras, “MotionHistory for Facial Action Detection in Video,” IEEE Conf. on Systems, Man and Cybernetics, Vol.1, pp.635–640, 2004.

[Viola01] Paul Viola, and Michael Jones, “Rapid object detection using a boosted cascade of

simple features," IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2001.

[Wren97] Christopher Richard Wren, Ali Azarbayejani, Trevor Darrell, and Alex Paul Pentland, "Pfinder: Real-Time Tracking of the Human Body," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.19, No.7, pp.780–785, 1997.

[Xia11] Lu Xia, Chia-Chih Chen, and J. K. Aggarwal, "Human Detection Using Depth Information by Kinect," Workshop on Human Activity Understanding from 3D Data in Conjunction with CVPR (HAU3D), pp.15-22, 2011.

[Yamaguchi10] 山口瑶子, 瀬々潤, "Web 閲覧履歴を用いた TV 番組推薦システム," DEIM Forum, A3-2, 2010.

[Yamakita06] 山北真実, 山田啓一, 山本修身, 山本新, "顔表情変化による携帯通話時の意識集中状態の検知," 信学技報, PRMU2006-7, pp.37-42, 2006

[Yamamoto04] 山本哲也，片渕典史，藪内勉，下倉健一朗に関する検討, "瞬目生起パターンに着目した注目度推定手法に関する検討," 電子情報通信学会技術研究報告, HIP, ヒューマン情報処理 104(168), pp.63-68, 2004.

[Yasuma11] Yuki Yasuma, and Miwa Nakanishi, "User characteristic-based information-providing service for museum with optical see-through head-mounted display," In Proc. of HCI (Human Computer Interaction), 2011

[Yoneya11] 米谷竜，川嶋宏彰，平山高嗣，松山隆司, "映像の顕著性変動と視線運動の時空間相関分析に基づいた集中状態推定," 情報処理学会研究会資料, CVIM178-16, 2011.

[Yoon01] Ho-Sub Yoon, Jung Soh, Younglae J. Bae, and Hyun Seung Yang, "Hand gesture recognition using combined features of location, angle and velocity," Pattern Recognition, vol.34, pp.1491–1501, 2001.

[Yu03] Xinguo Yu, Changshen Xu, Qi Tian, and Hon Wai Leong, "A ball tracking framework for broadcast soccer video," In Proc. of IEEE International Conference on Multimedia & Expo (ICME), Vol.II, pp.273–276, 2003.

研究業績

本論文を構成する論文

【学術論文】

1. **M. Takahashi**, M. Fujii, M. Naemura, S. Satoh，” Human Gesture Recognition System for TV Viewing using Time-of-Flight Camera,” Multimedia Tools and Applications, DOI 10.1007/s11042-011-0870-6, 2011.
2. **M. Takahashi**, M. Fujii, M. Shibata, S. Satoh，”Robust Recognition of Specific Human Behaviors in Crowded Surveillance Video Sequences,” EURASIP Journal on Advances in Signal Processing, vol. 2010, Article ID 801252, 14 pages, 2010. doi:10.1155/2010/801252.

【国際会議論文】（査読有）

3. **M. Takahashi**, M. Naemura, M. Fujii, S. Satoh，”Human Action Recognition in Crowded Surveillance Video Sequences by Using Features Taken from Key-Point Trajectories,” Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR 2011) Workshop on Machine Learning for Vision-based Motion Analysis (MLvMA 2011), pp.9-16, Colorado Springs, United States, July, 2011.
4. **M. Takahashi**, M. Fujii, M. Shibata, S. Satoh，”"Human gesture recognition using 3.5-dimensional trajectory features for hands-free user interface,” Proceedings of ACM International Conference on Multimedia (ACM Multimedia 2010) Workshop on Analysis and retrieval of tracked events and motion in imagery streams (ALTEMIS2010), pp. 3-9, Firenze, Italy, Oct, 2010.

【国際会議論文】（査読無）

（主著）

5. **M. Takahashi**, Y. Kawai, M. Naemura, M. Fujii, S. Satoh，'NHK STRL at TRECVID 2011: Surveillance Event Detection and High-Level Feature Extraction,” Proceedings of

TREC Video Retrieval Evaluation (TRECVID 2011 Workshop), Vol.1, 2011, p.267-275

6. **M. Takahashi**, Y. Kawai, M. Fujii, M. Shibata, N. Babaguchi, S. Satoh, ”"NHK STRL at TRECVID 2009: Surveillance Event Detection and High-Level Feature Extraction,” Proceedings of TREC Video Retrieval Evaluation (TRECVID 2009 Workshop), Vol.1, 2009, p.273-280

（共著）

7. Y. Kawai, **M. Takahashi**, M. Naemura, M. Fujii, S. Satoh, 'NHK STRL at TRECVID 2010: Semantic Indexing and Surveillance Event Detection,” Proceedings of TREC Video Retrieval Evaluation (TRECVID 2010 Workshop), Vol.1, 2010, p.304-310

【研究会】

8. 高橋正樹,藤井真人,苗村昌秀,佐藤真一, “特徴点軌跡に基づく監視映像からの人物行動検出,” 電子情報通信学会技術研究報告 PRMU, vol.110, no.414, PRMU 2010-225, 2011, p.111-116
9. 高橋正樹, 藤井真人, 苗村昌秀, 佐藤真一, “人物軌跡に基づく混雑映像からの特定行動検出,”電子情報通信学会技術研究報告 PRMU, vol.109, no.470, PRMU 2009-304, HIP2009-189, 2010. p.419-424

【全国大会】

10. 高橋正樹,藤井真人,苗村昌秀,佐藤真一, “ユーザの視聴状態とコンテンツ映像解析に基づく注目度推定,” 2011 年映像情報メディア学会冬季大会講演予稿集, 11-10, 2011
11. 高橋正樹,藤井真人,苗村昌秀,佐藤真一, “特徴点軌跡に基づく監視映像からの特定人物動作検出,” 2010 年映像情報メディア学会冬季大会講演予稿集, 6-8, 2010
12. 高橋正樹,藤井真人,苗村昌秀,佐藤真一, “3.5 次元時空間特徴量に基づく人物ジェスチャ認識手法”, 映像情報メディア学会年次大会予稿集, 4-6, 2010

学術論文（全て査読有）

【総研大入学以後の主著】

1. **M. Takahashi**, M. Fujii, M. Naemura, S. Satoh，” Human Gesture Recognition System for TV Viewing using Time-of-Flight Camera,” Multimedia Tools and Applications, DOI 10.1007/s11042-011-0870-6, 2011.
2. **M. Takahashi**, M. Fujii, M. Shibata, S. Satoh，”Robust Recognition of Specific Human Behaviors in Crowded Surveillance Video Sequences,” EURASIP Journal on Advances in Signal Processing, vol. 2010, Article ID 801252, 14 pages, 2010. doi:10.1155/2010/801252.
3. **M.Takahashi**, M.Fujii, M.Shibata, N.Yagi, S.Satoh, “Automatic pitch type recognition system from single-view video sequences of baseball broadcast videos,” International Journal of Multimedia Data Engineering and Management (IJMDEM), Vol.1, No.1, 2010, pp.12-36

【上記以外】

4. 高橋正樹，藤井真人，柴田正啓，八木伸行，“ゴルフ中継での放送カメラを用いたティーショット軌道表示システム,” 電子情報通信学会論文誌，D-2, vol.J92-D, no.7, 2009, p.1036-1044
5. 高橋正樹，三須俊枝，合志清一，藤田欣裕 “画像内の物体抽出技術を用いた高速投球軌跡作画手法,”電子情報通信学会論文誌，D-2 vol.J88-D-2, no.8, p.1681-1692, 2005
6. 金次保明，高橋正樹，三須俊枝，武智秀，加井謙二郎 “オブジェクト連動データ放送システムのメタデータの伝送手法とその評価実験”, 映像情報メディア学会誌 vol.60, no.5, p.781-788, 2006
7. 高橋正樹，三須俊枝，合志清一，藤田欣裕 “オブジェクト抽出技術のスポーツ番組への応用,”映像情報メディア学会誌 vol.59, no.1, p.159-165, 2005
8. 三須俊枝，高橋正樹，合志清一，蓼沼眞，藤田欣裕，八木伸行 “実時間画像処理に基づくオフサイドライン可視化システムの実用化”, 子情報通信学会論文誌，D-2 vol.J88-D-2, no.8, p.1681-1692, 2005
9. Y.Kanatugu, T.Misu, **M.Takahashi**, S.Gohshi, “The Development of an Object-linked Broadcasting System,”Journal of Broadcast Engineering, vol.9, no.2, p.102-109, 2004

国際会議・レター等

【総研大入学以後の主著】

（査読有）

1. **M.Takahashi**, M. Naemura, M. Fujii, S. Satoh, ”Human Action Recognition in Crowded Surveillance Video Sequences by Using Features Taken from Key-Point Trajectories,” Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR 2011) Workshop on Machine Learning for Vision-based Motion Analysis (MLvMA 2011), pp.9-16, Colorado Springs, United States, July, 2011.
2. **M.Takahashi**, M. Fujii, M. Shibata, S. Satoh, ”'Human gesture recognition using 3.5-dimensional trajectory features for hands-free user interface,” Proceedings of ACM International Conference on Multimedia (ACM Multimedia 2010) Workshop on Analysis and retrieval of tracked events and motion in imagery streams (ALTEMIS2010), pp. 3-9, Firenze, Italy, Oct, 2010.

（査読なし）

3. **M.Takahashi**, Y. Kawai, M. Naemura, M. Fujii, S. Satoh, 'NHK STRL at TRECVID 2011: Surveillance Event Detection and High-Level Feature Extraction,” Proceedings of TREC Video Retrieval Evaluation (TRECVID 2011 Workshop), Vol.1, 2011, p.267-275
4. **M.Takahashi**, Y. Kawai, M. Fujii, M. Shibata, N. Babaguchi, S. Satoh, ”'NHK STRL at TRECVID 2009: Surveillance Event Detection and High-Level Feature Extraction,” Proceedings of TREC Video Retrieval Evaluation (TRECVID 2009 Workshop), Vol.1, 2009, p.273-280

【上記以外】

（主著・査読有）

1. **M.Takahashi**, M.Fujii, N.Yagi, “Automatic pitch type recognition from baseball broadcast videos,”ISM2008(The IEEE International Symposium on Multimedia 2008), pp.15-22, 2008
2. **M.Takahashi**, T.Misu, M.Naemura, M.Fujii, N.Yagi,“Enrichment system for live sports broadcasts using real-time motion analysis and computer graphics,” BroadcastAsia 2007 International Conference, 2007

3. **M.Takahashi**, T.Misu, M.Tadenuma, N.Yagi,“Real-time ball trajectory visualization using object extraction” IEE 2nd European Conference on Visual Media Production (CVMP) No.008, pp.62-69, 2005

（共著・査読有）

4. M.Naemura, **M.Takahashi**, M.Fujii, N.Yagi, “A Method of Multi-factorization for Recognizing Emotions from Gestures,” 8th IEEE International Conference on Automatic Face and Gesture Recognition (FG2008), poster session 2
5. 三須俊枝，**高橋正樹**，藤井真人，八木伸行 “パーティクルフィルタによる単眼動画像からのサッカーボール 3 次元軌道推定”, 第 5 回情報科学技術フォーラム情報科学技術レターズ, vol.5, LI-002, p.167-170, 2006
6. 三須俊枝，**高橋正樹**，蓼沼眞，八木伸行 “サッカー映像のフォーメーション解析に基づく実時間イベント検出”第 4 回情報科学技術フォーラム情報科学技術レターズ, vol.4, LI-003, p.141-144, 2005
7. 三須俊枝，小田原邦治，**高橋正樹**，合志清一，宮越肇，藤田欣裕 “画像オブジェクト追跡に基づくサッカーオフサイドラインの可視化”, 第 3 回情報科学技術フォーラム情報科学技術レターズ vol.3, LI-009, p.187-190, 2004
8. Y.Kanatsugu, T.Misu, **M.Takahashi**, S.Gohshi,“The Development of an Object-linked Broadcasting System,”International Workshop on Advanced Image Technology 2004(IWAIT 2004), p.401-406, 2004
9. 三須俊枝，苗村昌秀，**高橋正樹**，和泉吉則 “オブジェクト追跡と背番号認識の連携による動画像用スポーツ選手同定手法”, 第 2 回情報科学技術フォーラム情報技術レターズ vol.2, LI-012, p.187-189, 2003

（共著・査読なし）

10. Y. Kawai, **M.Takahashi**, M. Naemura, M. Fujii, S. Satoh, 'NHK STRL at TRECVID 2010: Semantic Indexing and Surveillance Event Detection,” Proceedings of TREC Video Retrieval Evaluation (TRECVID 2010 Workshop), Vol.1, 2010, p.304-310
11. Y. Kawai, **M.Takahashi**, M. Sano, M. Fujii, M. Shibata, N. Yagi, N.Babaguchi, “NHK STRL at TRECVID 2008: High-Level Feature Extraction and Surveillance Event Detection,” Proceedings of TREC Video Retrieval Evaluation (TRECVID 2008 Workshop), Vol.1, 2008, p.358-365

学術講演・研究報告等

○ 研究会（主著）

1. **高橋正樹**，三須俊枝，合志清一 “オブジェクト抽出技術のスポーツ番組への応用” 電子情報通信学会技術研究報告 IE vol.103, no.383, IE2003-68, DSP2003-108, ICD2003-106, p.1-6, 2003
2. **高橋正樹**，三須俊枝，合志清一，藤田欣裕 “野球の投球軌跡表示手法”映像情報メディア学会技術報告 vol.28, no.61, ME2004-163, p.25-28, 2004
3. **高橋正樹**, 三須俊枝, 蓼沼眞, 三角和浩, 八木伸行 “野球投球軌跡作画装置“B-Motion” の開発・運用報告”電子情報通信学会技術研究報告 ITS vol.104, no.647, ITS2004-86, IE2004-220, p.119-124, 2005
4. **高橋正樹**，三須俊枝，藤井真人，八木伸行 “移動体の抽出および動き予測に基づくゴルフのティーショット軌道表示システム”, 映像情報メディア学会技術報告 vol.30, no.41, AIT2006-107, p.17-20, 2006
5. **高橋正樹**，藤井真人，八木伸行 “投球軌道と捕手の動作特徴に基づく野球の球種識別”電子情報通信学会技術研究報告 PRMU vol.107, no.427, PRMU2007-162, p.29-34, 2008
6. **高橋正樹**, 藤井真人, 苗村昌秀, 佐藤真一, “人物軌跡に基づく混雑映像からの特定行動検出,”電子情報通信学会技術研究報告 PRMU, vol.109, no.470, PRMU 2009-304, HIP2009-189, 2010. p.419-424
7. **高橋正樹**,藤井真人,苗村昌秀,佐藤真一, “特徴点軌跡に基づく監視映像からの人物行動検出,” 電子情報通信学会技術研究報告 PRMU, vol.110, no.414, PRMU 2010-225, 2011, p.111-116

○全国大会（主著）

8. **高橋正樹**，片山美和，冨山仁博，岩舘祐一 “インタラクティブ操作が可能な三次元映像オブジェクト用ブラウザの開発”, 2003 年映像情報メディア学会年次大会講演予稿集 1-6, 2003
9. **高橋正樹**，三須俊枝，合志清一，藤田欣裕 “高速移動オブジェクトの抽出・追跡法に関する一考察”2003 年映像情報メディア学会冬季大会講演予稿集 8-3, p.83, 2003

10. **高橋正樹**，三須俊枝，小田原邦治，三角和浩，合志清一，西沢利勝，藤田欣裕 “オブジェクト抽出技術による投球軌跡作画システムの開発”, 2004 年映像情報メディア学会年次大会講演予稿集 7-4, 2004
11. **高橋正樹**，三須俊枝，藤井真人，八木伸行“ゴルフ中継におけるティーショット軌道表示システム”2006 年電子情報通信学会総合大会講演論文集，情報・システム 2 D-12-38, p.170, 2006
12. **高橋正樹**，三須俊枝，藤井真人，八木伸行 “野球中継番組での投球軌跡表示”第 6 回日本プラント・ヒューマンファクター学会総会 ヒューマンファクターズ特別号, p.35-36, 2006
13. **高橋正樹**，小田健市，三須俊枝，苗村昌秀，藤井真人，八木伸行 “ゴルフパッティングのリアルタイム軌跡表示,”精密工学会 動的画像処理実利用化ワークショップ 2007, I2-8, p.260-263, 2007
14. **高橋正樹**，三須俊枝，苗村昌秀，藤井真人，八木伸行, “飛翔するゴルフボール抽出処理の全自動化”2007 年電子情報通信学会総合大会講演論文集，情報・システム 2, D-12-44, p.160, 2007
15. **高橋正樹**，苗村昌秀，藤井真人，八木伸行 “B-Motion の軌跡データに基づく野球の球種識別手法”第 6 回情報科学技術フォーラム一般講演論文集, No.3, H-063, p.153-154, 2007
16. **高橋正樹**，苗村昌秀，藤井真人，八木伸行 “パーティクルフィルタを用いたティーショットシーンでのゴルフボール追跡”, 第 6 回情報科学技術フォーラム一般講演論文集, No.3, H-064, p.155-156, 2007
17. **高橋正樹**，藤井真人，八木伸行 “ゴルフ中継番組におけるボール軌道の可視化” 第 4 回デジタルコンテンツシンポジウム講演予稿集 1-4, 2008
18. **高橋正樹**，藤井真人，柴田正啓，八木伸行 “オプティカルフローと機械学習を用いた映像中の特定動作検出”第 3 回日本プラント・ヒューマンファクター学会ポスターセッション予稿集, p.34-35, 2008
19. **高橋正樹**，藤井真人，柴田正啓，八木伸行 “ボウリング中継番組におけるボール軌道の可視化”映像情報メディア学会冬季大会講演予稿集, 4-2, 2008
20. **高橋正樹**，藤井真人，柴田正啓，八木伸行 “ボウリング中継番組におけるボール軌道作画装置の運用,” 第 8 回情報科学技術フォーラム講演論文集(FIT 2009), no.3, H-045, 2009, p.199-200
21. **高橋正樹**,藤井真人,苗村昌秀,佐藤真一, “3.5 次元時空間特徴量に基づく人物ジェスチャ認識手法”, 映像情報メディア学会年次大会予稿集, 4-6, 2010
22. **高橋正樹**,藤井真人,苗村昌秀,佐藤真一, “特徴点軌跡に基づく監視映像からの特定

人物動作検出," 2010 年映像情報メディア学会冬季大会講演予稿集, 6-8, 2010

23. 高橋正樹,藤井真人,苗村昌秀,佐藤真一, "ユーザの視聴状態とコンテンツ映像解析に基づく注目度推定," 2011 年映像情報メディア学会冬季大会講演予稿集, 11-10, 2011

○研究会（共著）

24. 苗村昌秀，高橋正樹，三須俊枝，和泉吉則 "Hough パラメータ群の一致計算による動き推定手法"映像情報メディア学会技術報告 vol.27, no.8, HIR2003-17, ME2003-17, AIT2003-17, p.91-96, 2003
25. 金次保明，三須俊枝，高橋正樹，苗村昌秀 "オブジェクト連動データ放送システムの開発とその記述方式"情報処理学会研究報告 AVM vol.2003, no.81, 2003-AVM-41(4), p.17-22, 2003
26. 金次保明，三須俊枝，高橋正樹，合志清一，苗村昌秀 "オブジェクト連動データ放送システムの開発"映像情報メディア学会技術報告 vol.27, no.62, BCT2003-43, p.1-4, 2003
27. 冨山仁博，岩舘祐一，片山美和，高橋正樹，今泉浩幸"多視点画像を用いた 3 次元映像の高精細生成システム,"情報処理学会研究報告 CVIM, vol.2003, no.88, pp.57-62, 2003
28. 金次保明，三須俊枝，高橋正樹，合志清一，苗村昌秀"オブジェクト連動データ放送システムの開発,"映像情報メディア学会研究報告，vol.27, no.62, pp.1-4, 2003
29. 三須俊枝，高橋正樹，藤井真人，八木伸行"スポーツ番組におけるメタデータ自動生成 ～ サッカー選手追跡・同定のためのデータフュージョン ～," 電子情報通信学会技術研究報告 PRMU，vol.105, no.415, p.39-44, 2005
30. 三須俊枝，高橋正樹，藤井真人，八木伸行"スポーツ戦術実況のための実時間画像解析システムの開発," 精密工学会 動的画像処理実利用化ワークショップ, 7.2, pp.233-238, 2006
31. 宮崎勝，藤沢寛，木村徹，西村敏，浜口斉周，大竹剛，望月貴裕，高橋正樹，米倉律，東山一郎，小川浩司，井田美恵子"次世代公共放送サービスモデルの開発と実験," 電気四学会関西支部専門講習会「多様化するメディアの現状と動向」
32. 山内結子，高橋正樹，奥田誠，三ッ峰秀樹，サイモン クリピングデル，苗村昌秀, 藤井真人 "テレビユーザインタフェースに向けた視聴状況調査に関する一考察," 電子情報通信学会技術研究報告 MVE, vol.110, no.475, IE2010-149, MVE2010-137, 2011, p.19-24
33. 苗村昌秀，高橋正樹，山内結子，藤井真人"CRF を用いた TV の興味視聴区間の

推定手法,”電子情報通信学会研究報告 PRMU, 2012

○全国大会（共著）

34. 冨山仁博，片山美和，岩舘祐一，**高橋正樹**“IEEE1394 カメラを用いた取り囲み型 3 次元映像生成システムの開発,”映像情報メディア学会年次大会講演予稿集 2-7, 2003
35. 三須俊枝，**高橋正樹**，合志清一，藤田欣裕“投球軌跡作画システムにおける欠落データの適応的内外挿手法,” 映像情報メディア学会年次大会講演予稿集, 7-5, 2004
36. 三須俊枝，**高橋正樹**，藤井真人，八木伸行“逐次モンテカルロ法による実時間サッカーボール追跡,” 電子情報通信学会総合大会講演論文集, D-12-39, pp.171, 2006
37. 苗村昌秀，**高橋正樹**，藤井真人，八木伸行“多重分解処理を用いたジェスチャからの感情認識,” 情報科学技術フォーラム(FIT)一般講演論文集，no.3，H-049, pp.115-116, 2007
38. 住吉秀樹，柴田正啓，藤井真人，後藤淳，山田一郎，望月貴裕，松井淳，三須俊枝，宮崎勝，**高橋正樹**，河合吉彦，三浦菊佳，八木伸行“CurioView：情報検索を活用した新しい視聴スタイルの提案,” 映像情報メディア学会年次大会講演予稿集, 7-5, 2008
39. 柴田正啓，後藤淳，山田一郎，望月貴裕，松井淳，三須俊枝，宮崎勝，**高橋正樹**，河合吉彦，三浦菊佳，住吉秀樹，藤井真人，八木伸行“検索技術を使う新しいテレビ視聴スタイル　CurioView,” 情報科学技術フォーラム(FIT)講演論文集，no.3, H-007, pp.77-78, 2008
40. 佐野雅規，住吉秀樹，後藤淳，望月貴裕，宮崎勝，三浦菊佳，河合吉彦，**高橋正樹**，三須俊枝，松井淳，サイモンクリピングデル，藤井真人，柴田正啓，八木伸行“番組を推薦するテレビ　CurioView,” 情報科学技術フォーラム(FIT)講演論文集, no.3, K-049, pp.647-648, 2009
41. 浜口斉周，藤沢寛，宮崎勝，西村敏，木村徹，大竹剛，望月貴裕，**高橋正樹**，米倉律，小川浩司，東山一郎“VOD と番組レビューSNS を組み合わせたサービスにおける視聴行動の変化,” 秋季経営情報学会全国研究発表大会予稿集, E1-2, 2010
42. 宮崎勝，藤沢寛，木村徹，西村敏，浜口斉周，大竹剛，望月貴裕，**高橋正樹**，米倉律，小川浩司，東山一郎“SNS を利用した番組視聴機会拡大に関する検討 -番組レビューサイト Teleda による実証実験-,” 映像情報メディア学会冬季大会講演予稿集, 5-5, 2010
43. 宮崎勝，浜口斉周，**高橋正樹**，木村徹，西村敏，大竹剛，有安香子，望月貴裕，

藤沢寛，”番組レビューSNS サイト“teleda”におけるユーザの視聴行動に関する検証,” 2011 年映像情報メディア学会年次大会講演予稿集, 1-8, 2011

44. 三ッ峰秀樹，苗村昌秀，サイモンクリピングデル，山内結子，**高橋正樹**，奥田誠，藤井真人“状況理解に基づくテレビユーザーインターフェースの提案,” 映像情報メディア学会冬季大会講演予稿集, 5-3, 2010
45. 山内結子，奥田誠，**高橋正樹**，サイモン　クリピングデル，苗村昌秀，藤井真人“テレビ視聴インターフェイス-UTAN-の提案,” 映像情報メディア学会冬季大会講演予稿集，2011

解説記事等

（主著）

1. **高橋正樹**，三須俊枝，合志清一，藤田欣裕 “ボール距離情報のリアルタイム可視化装置,”画像ラボ vol.16, no.6, pp.29-33, 2005
2. **高橋正樹**，“「考える」ＴＶ,”電子情報通信学会誌，vol.92, no.1, 2009, p.25
3. 高橋正樹，“映像オブジェクトの軌跡画像合成装置，映像オブジェクトの軌跡画像表示装置およびそのプログラム,”VIEW, vol.28, no.1, pp.10, 2009
4. **高橋正樹**，藤井真人，柴田正啓，八木伸行，“放送カメラを用いたティーショット軌道表示システム,”画像ラボ, vol.21, no.3, pp.27-32, 2010
5. **高橋正樹**, “画像認識技術のスポーツ番組への応用,” 技術月刊誌 OplusE, vol.32, no.6, 2010, p.704-706
6. **高橋正樹**, “TRECVID 2009,”情報処理学会誌, vol.51, no.8, 2010, pp.1068-1073
7. **高橋正樹**, 国際会議“ACM Multimedia2010,” 電子情報通信学会, vol.94, no.4, 2011, p.348
8. **高橋正樹**，佐野雅規“ACM Multimedia2011 レポート,”映像情報メディア学会誌, vol.66, no.4, 2012

（共著）

9. 三須俊枝，**高橋正樹**，藤井真人，八木伸行 “スポーツ戦術実況のための実時間画像解析システムの開発,”画像ラボ vol.18, no.1, pp.38-43, 2007
10. 小田健市，**高橋正樹** “ゴルフ中継のパット軌跡表示,”放送技術 vol.60, no.1, pp.60-63, 2007
11. 苗村昌秀，**高橋正樹**，奥田誠，三ッ峰秀樹，サイモンクリピングデル，藤井真人“視聴状況に基づいたテレビインターフェース,”映像情報メディア学会誌，vol.64, no.12, pp.1816-1819, 2010

登録特許

1. 特許第 4181473 号，映像オブジェクト軌跡合成装置，その方法及びそのプログラム
2. 特許第 4268479 号，距離情報付加装置，付加映像生成装置，付加映像生成方法及び距離情報付加プログラム
3. 特許第 4441354 号，映像オブジェクト抽出装置，映像オブジェクト軌跡合成装置，その方法及びそのプログラム
4. 特許第 4555690 号，軌跡付加映像生成装置及び軌跡付加映像生成プログラム
5. 特許第 4546810 号，軌跡付加映像生成装置及び軌跡付加映像生成プログラム
6. 特許第 4695615 号，映像オブジェクトの軌跡画像合成装置およびそのプログラム
7. 特許第 4728795 号，人物オブジェクト判定装置及び人物オブジェクト判定プログラム
8. 特許第 4758842 号，映像オブジェクトの軌跡画像合成装置，映像オブジェクトの軌跡画像表示装置およびそのプログラム
9. 特許第 4881178 号，走行距離映像生成装置及び走行距離映像生成プログラム
10. 特許第 4886707 号，オブジェクト軌道識別装置，オブジェクト軌道識別方法，及びオブジェクト軌道識別プログラム
11. 特許 4906588 号，特定動作判定装置、リファレンスデータ生成装置、特定動作判定プログラムおよびリファレンスデータ生成プログラム

表彰等

1. 映像情報メディア学会　技術振興賞（放送番組技術賞）　2005 年 6 月 4 日
2. 映像情報メディア学会　船井賞（技術賞）　2005 年 6 月 4 日
3. 電子情報通信学会　学術奨励賞　2008 年 3 月 19 日
4. 情報処理学会　FIT ヤングリサーチャー賞　2008 年 9 月 3 日
5. 映像情報メディア学会　デジタルコンテンツシンポジウム 船井賞　2009 年 6 月