論文の要旨

青木　恒

人間が自己の置かれた環境を認知したり，他者による意図表現を理解したりするのに際して視覚の果たす役割は大きい．このためコンピュータに人間行動の支援を行わせようとするアプローチの一つとして映像認識に関する研究が着目されてきた．一方で，近年ではハードウェアの高集積化，低価格化が進んだことにより，小型で安価なビデオカメラ，ビデオデッキが急速に普及し，最近では HDD/DVD レコーダの登場やテレビ視聴・録画が可能なパーソナルコンピュータ(PC)の登場など，家庭においても映像記録のデジタル化が進んでおり，機器の普及とあいまって記録容量の拡大もめざましい．

このように様々な映像記録が日常生活の中で可能になり，また，記録されるデジタル映像の量が増大するに伴って，構造化，関連付け，アノテーション(記述情報やメタデータの付与)の必要性が高まっている．一つの側面としては，人間の視点と同様に映像を撮影するカメラを身に付けたウェアラブル・コンピューティングの世界において，映像に対して計算処理を行うことによって，ユーザの居場所や対話の相手などを認識することで状況理解を行い，リアルタイムでユーザが必要としている情報を提示したり，あるいは生活記録映像を後刻認識処理することによって自動日記を作成し，記憶の想起や今後の行動の指針として活用したりするようなアプリケーションが考えられる．本研究では人間の日常生活の映像を実世界映像と称し，ユーザの場所や対話相手といった付加情報で実世界映像をインデクシングすることにより，人間の行動をナビゲーションすることを目指した技術の開発に取り組んだ．

映像構造化が必要となるもう一つの側面としては，テレビ番組などの表現映像が大量に家庭でデジタル録画される時代の到来に伴い，ユーザが所望の場面へのアクセスが困難になりつつある問題を解決しなければならないという課題が挙げられる．本研究ではこうした表現映像をコンテンツ映像と称し，映像が持つ場面やコーナー，話題といった構造を推定するインデクシングを行うことにより，コンテンツ映像へのアクセスをナビゲーションすることを目指した技術の開発にも取り組んだ．

上記のように本研究では人間に入力される映像情報のうち，テレビ・ビデオ等を視聴している時間に関する「コンテンツ映像」のインデクシング技術と，それ以外の，実世界に生活している時間に関する「実世界映像」のインデクシング技術の双方を開発することにより，ユーザサイドにおいて適切なナビゲーションを実現するための基礎となる技術のラインナップをこの両面から構築した．ここでは処理速度をいかに軽量化するかという視点も実用面から極めて重要な問題となる．実世界映像のインデクシングに関しては比較的低い

計算能力を持ったポータブル/ウェアラブル･コンピュータ上での実装が必要となり,また，即時的なアドバイス提示という目的の達成のためにはリアルタイムで処理結果を出力する必要がある．一方，コンテンツ映像に関しては PC だけではなく，近年の急速な普及が目覚しい HDD/DVD レコーダにおいても実現可能でなければより多くのユーザに技術の価値を提供できない．加えて，録画コンテンツの増大に伴ってインデクシング処理に要する時間の増大を招いてしまうと，ユーザに利便性を提供するためのインデクシング処理であるにもかかわらず，処理終了までの待機時間が増大することにより，必ずしも利便性提供と整合しない機能提供となる恐れがある．このため，実世界映像インデクシング，コンテンツ映像インデクシングのいずれにおいても，それが搭載されると想定されるプラットフォーム上で極力リアルタイムで映像処理が完了することが実用的なナビゲーション機能の提供のためには必須条件となる．

本論文では上記のように，日常生活行動や記憶想起のためのナビゲーションを行うための基礎技術として実世界映像のインデクシング，大量の番組録画映像から所望の場面に容易にアクセスできるようなナビゲーションを行うための基礎技術としてコンテンツ映像のインデクシングに取り組みつつ，実用的な軽量処理でそれらを実現するためのアルゴリズム提案および検証結果について議論する．

研究の背景について詳細に述べた第 1 章，映像ナビゲーションのためのインデクシング技術についての動向をまとめた第 2 章に引き続き，第 3 章では身につけたカメラの画像から次元数の少ない特徴量を抽出し，特徴量の時系列変化を辞書として持つデータベースと DP マッチングによってリアルタイムで比較することによって，ユーザの現在地を検出する技術について論じる.本論文の手法では,ナビゲーションを受けたいユーザ自身が所持(装着)したカメラからの映像のみを用いて位置の認識を行うため，周辺環境側にはタグ等を設置する必要がない．このため，タグの設置やメンテナンスの手間やコストが制約となることがない．また，認識システム側のハードウェア構成もカメラと計算装置(PC)のみという最低限のもので実現できる.処理アルゴリズムも 166MHz という比較的低速のプロセッサで処理できる程度の軽量であり，これは現在では PDA や，組み込み系プロセッサを搭載した多くのポータブル装置などが持つ処理性能あるいはそれ以下の処理性能で実現できることを意味している．この手法は，GPS などで十分な空間解像度が得られない室内環境でも用いることができ，環境光の変化に対してもロバストであるため，ウェアラブル・コンピュータへの応用などが考えられる．ユーザの居場所に応じたナビゲーションや自動日記へのインデックスデータとして用いることも可能にするものである．

第 4 章では赤外線 LED を配した小型のアクティブタグをユーザが身につけ，ポータブルな映像処理機器によってアクティブタグの発する ID 情報を読みとる研究について論じる.本

章では人間への影響が極めて少ない近赤外線を発信するビジュアルタグをカメラで読み取るという方法により，既存のタグシステムである無線タグとバーコードの短所を相補的に克服する方法を提案，タグの設計および認識アルゴリズムの実装を行い，手法の有効性を実証する．処理装置は 600MHz という，現在ではミニノート PC 搭載のプロセッサとしては最軽量級のハードウェアで実現できる．タグについても軽量なものであり，人間には通常の名札のように見せながら，赤外線光源を秘匿して内蔵することも可能である．本手法により，正対している人物に関する名前や所属，関心などの情報を撮影画像にリアルタイムでスーパーインポーズするシステムを試作した．このインタフェースを発展させることにより，前回その人物といつあったか，あるいはその人物と話すべき話題といった情報を提示できるため，相手に対応したスムーズな会話を促したり，予定していた重要なインタラクションをしそこなうのを防いだりといった，行動ナビゲーションを行うことができる．また，本システムの特長として ID 発信者とユーザとの位置関係が計算できるため，複数の人物に正対しているときにも，人物と情報との対応付けが明確であり，人物を弁別して適切な会話を斡旋することができるといったナビゲーションを可能にするものである．

上記が実世界映像のインデクシングに関するものであるのに対し，以下 2 つの章ではコンテンツ映像のインデクシング技術について論じる．第 5 章で論じる研究では，試行として映画を対象とし，映像コンテンツ中に繰り返し登場する類似ショットの登場パターンを利用することによって番組構成を一定の意味的単位に分割する手法を確立した．従来のカット検出のみによるショットサムネイル一覧ではサムネイル数が莫大になるという問題を解決するため，類似ショットからの重複サムネイルの省略，意味的単位に即したサムネイル表示など，インデクシング結果を利用した 3 通りの内容一覧表示が可能な映像ブラウジングシステムを試作し，サムネイル数の抑制効果について実験，検証した．本手法ではカット検出のみによる映像分割と比べてセグメント数を最大で 7 分の 1 に縮約させることができた.また,類似ショットが真に類似であるかどうかを検証する確信度の計算法を導入し，類似ショットの誤検出(過検出)を 38%削減するなど，コンテンツ映像インデクシングによるナビゲーション実現の基礎となる技術である．

第 6 章では，第 5 章の成果をさらに発展させ，類似ショットの登場パターンから番組の構成上，関連が密接である「対話部分」を定量的に発見する指標を定義し，これを用いることによってニュース番組やバラエティ番組に対する軽量で高速なインデクシングを可能にするアルゴリズムについて議論する．提案手法では原映像の 2,300 倍程度も小さい特徴量の計算処理のみですむため， 733MHz の CPU を用いた場合，リアルタイム処理の場合，特徴量の計算と類似ショットの検出までに要する CPU 使用率は 0.3〜0.4%程度となり，類似ショット検出結果を用いたセグメント作成までの所要時間は測定限界付近の早さであった．理想のセグメント数に対する検出セグメント数で定量的に評価し，適合率においては本手

法の導入によってカット検出のみのセグメンテーションと比較して 9〜196 倍正解率が向上した.このことから組み込み系プロセッサを搭載したホームサーバや HDD/DVD レコーダ，あるいは映像処理専用処理チップ/ボードなど，より普及している安価な製品へ搭載され，コンテンツへのアクセス・ナビゲーションの実現をも視野に入る処理速度を実現した.

最後に第 7 章では，上記のような本研究全体を通じて得られた成果やその考察，今後の課題について議論する.