論文の内容の要旨

論文題目　「索引付けのための映像とテキスト教材の対応付けシステム」

氏名　　濵田　玲子

・背景と目的

近年、テレビやビデオ、WWW などを通して発信される膨大なマルチメディアデータを収集・整理し、効率の良い利用法を模索する研究が盛んに進められている。マルチメディアデータとは、主に画像・音声・テキスト情報が同期したデータ群である。従来は画像を利用した解析が主流であったが、画像認識単独での映像の意味の把握や、高度な構造化は非常に困難であった。そこで 1990 年代に入り、各メディアからの情報を統合することによってより簡単な処理でより大きな効果を得ようとする統合処理が検討されるようになった。各メディアにおける要素技術は、歴史が古いこともあり数多く開発されているが、統合そのものに関してはいまだに単純な方法を採用している研究も多い。

そこで我々は、完全には同期していない複数のメディアを統合的に処理することにより、実用的な統合システムの検討・構築を目指している。本研究では、メディアの中でも比較的意味情報を抽出しやすいテキストに着目し、テキスト教材の存在する教養番組の映像とテキスト教材の対応づけを目指す。複数メディアからの情報を有効に統合するため、テキストの解析結果を映像理解に反映させることで、各メディア処理単体での困難な点を回避し、より効果的な処理を目指す。

映像の索引付けに関してはこれまでにも様々な研究がなされており、対象を一般化しようとする研究も多い。しかし、個々の要素技術の限界、また対象に固有の知識を利用できないことなどにより、対象を限定した場合よりも精度が低下することがほとんどである。さらに、索引の種類が一般的なものに限られるため、対象映像の種類によっては処理自体が有効でなくなる。そこで本研究では、対象を限定してその特徴を利用することにより、より高精度かつ効果的な索引付けを行なう。

ここで、本研究では教養番組の中でも最も親しまれている料理番組に対象を絞ったシステムを提案する。料理番組にはほとんどの場合テキスト教材が存在するが、教材では表現しきれない様々な情報が映像に含まれており、テキストと映像の情報を統合的に利用することの効果は大きい。本研究では対象を料理番組に限定することで、対象に関する知識を最大限に活かした実用的な統合システムの構築を目指している。最終的には、料理番組にお

ける映像とテキスト教材を対応付けることで、映像とテキストの対応する各部分がリンクされた新しい形態のマルチメディアデータの自動生成を目標とする。これにより、抽象度の高い適切な索引が映像につき、台所環境における調理支援システムやマルチメディアデータベースの構築およびその検索など、統合されたマルチメディアデータを利用した様々な応用アプリケーションの開発も可能となる。

・提案システムの概要

本研究では、テキスト教材の情報を最大限に利用するため、まずテキスト教材における調理手順の構造解析を行なう。その際に、大量のテキスト文書を解析して作成した独自の辞書を利用する。映像処理においては、まずカット検出およびショット分類を行ない、さらに動き及び背景の解析を行なって、映像構造を抽出する。音声からは、クローズドキャプションを利用したテキスト処理によってキーワード抽出を行なう。最後に、映像および音声情報と統合することによりテキストから抽出された手順構造の制約条件を解き、映像とテキストの対応付けを行なう。

まずテキスト処理部においては、料理テキスト教材における調理手順の説明文書に対して構造解析を行ない、手順のフローグラフを抽出する。この際には、大量のテキスト文書から統計的に抽出して手動で訂正した辞書を利用し、文脈解析を行なう。調理手順における構造は、複数の素材が調理されたり混ぜられたりして最終的に一つの料理としてまとまるツリー構造のフローグラフとなる。このフローグラフから調理順序の制約条件を抽出することで、映像における調理手順との対応付けに利用する。

次に、映像の区切り検出における最も重要なヒントはカット点である。本研究においては、DCT クラスタリングを利用するカット検出手法を導入し、映像をショットに分割する。料理映像におけるショットは大きく手元ショットおよび人物ショットに分けられる。我々は、肌色の統計データおよび料理映像の特徴を利用した顔認識手法によって、人物ショットと手元ショットの自動分類を 90%以上の精度で実現した。料理映像においては人物ショットと手元ショットがほぼ交互に出現する。視覚的には動作や道具などが大映しにされる手元ショットが特に重要であると考えられるが、その中にも特に重要な部分と、動作と動作の間など比較的冗長な部分が含まれる。そこで我々はこのように手元ショットのなかにさらに含まれる構造を解析し、各手元ショットの特徴を抽出することで、テキストとの統合処理の際に有効な情報として利用する。

ここで、料理映像においては特に調理動作に関する視覚的情報が重要であると考えられる。そこで、画面全体の動きの大きさを解析することで、動きによる映像構成の推測を行なう。さらに、全体的な動き解析のみでは動きの種類や特に重要な動作などを区別することでき

ないため、繰り返し動作の自動検出手法を提案する。この手法においては、特に料理映像においては重要な動作の多くが繰り返し動作であることに着目し、その周期性を利用することで、約90%の精度で繰り返し動作の自動検出を実現した。

また、料理映像の構造は、動作の有無の他に、各動作が行なわれている背景から分析することができる。料理映像においてはほとんどの調理はすべて台所で行なわれるが、同じ台所でも、レンジ台、流し台、調理台など、動作の特徴によって背景が異なることが多い。従って、背景を解析することによって動作のおおまかな種類やその順序などの情報を得ることができる。本研究では、あらかじめ教師つき学習によって複数の料理番組に共通の画面構成を抽出し、画面内で背景が映る確率の高い位置を特定した。これにより、色情報によるクラスタリングによって高精度な背景の自動分離を実現した。

最後に、映像中の音声内容も統合処理の際には大きなヒントとなる。本研究では、テレビ局から提供されるクローズドキャプションに言語処理を施してキーワード抽出を行ない、その結果を統合処理に利用する。

統合処理部においては、テキスト処理によって抽出されたテキストのツリー構造と、映像における線形な順序構造を対応づける必要がある。そのため、映像における背景の構成、またクローズドキャプションからの音声内容に関する情報を利用し、テキスト教材の順序構造を解く。順序構造を解くのと同時に、映像にテキストの各部分が対応づけられる。動きの解析結果を利用することで、動作単位の細かい索引まで映像につけることが可能となる。

・むすび

本研究では、料理映像を題材としたマルチメディア統合処理システムの構築を行なった。これにより、対象に関する知識およびテキスト情報を活用した統合処理によって、有用な索引付けシステムの構築が可能であることを示した。

本システムの手法を応用することで、実験・組み立てなど、手順書つきのインストラクション・ビデオに対する索引付けが可能になると考えられる。特に料理映像は教育的な内容である上に生活に密着しているため、様々な実用的な応用が可能となる。例えば料理映像の自動要約による閲覧システム、マルチメディアデータベース、マルチメディア調理支援・教育システムなどが挙げられる。また、適切なセンサなどと組み合わせれば、インテリジェント・キッチン、自動調理システムなどへの応用も考えられる。