審 査 の 結 果 の 要 旨

氏　　名　　柴田　知秀

本論文は、「Structural Understanding of Instruction Videos by Integrating Linguistic and Visual Information」（言語情報と映像情報の統合による作業教示映像の構造的理解）と題し、作業教示映像の高度利用を目的とし、言語情報と映像情報を統合することによりその内容の構造的理解を行ない、実験によりその有効性を論じたものであり、６章から構成されている。

第１章は「Introduction」（緒言）であり、実世界情報処理を行なうためのメディア処理統合の必要性を述べ、本研究で扱う映像処理でこれまで利用されてきた言語処理を概観している。これに対し、本研究のアプローチである構造的言語処理とそれに基づく言語と映像の統合処理について述べている。

第２章は「Linguistic Analysis of Instruction Utterances」（作業教示発話の言語処理）と題し、料理教示発話を具体的題材とし、作業教示発話（クローズドキャプション）の言語解析について述べている。まず、Webテキストから自動構築した大規模格フレームに基づく省略解析を行なう。省略解析は以下で述べる談話構造解析や言語と映像との対応付けをとる際に重要となる処理である。次に、発話タイプの認識を行なう。作業教示発話の場合、作業に関する発話が中心となり、今の料理の状態を示す発話や作業の留意事項などといった発話が補足的な役割を果たす。これらのタイプを正確に認識することは、重要発話抽出や、以下で述べるトピック推定において重要となる。そこで、節を基本単位とし、発話タイプの認識を行なう。そして、省略解析結果・発話タイプ認識結果・接続詞などの情報を利用し、文または節間の関係を明らかにし、局所的な修飾関係を解析する。節間の関係は、節末が「～て」なら「順接」といった節末のパターンで認識し、文間の関係は、接続詞「なぜなら」で始まる文は前文と「理由」の関係になりやすいことや、発話タイプが「留意事項」の文は前文と「詳細化」の関係になりやすいといったことを考慮する。これらの情報を考慮することにより、談話関係を高精度に認識できることを明らかにした。

第３章は「Unsupervised Topic Identification based on HMMs」（隠れマルコフモデルに基づく教師なしトピック推定手法）と題し、言語情報と映像情報を用いて、教師なし学習でトピック(下ごしらえ、炒める、盛り付けなど)の推定を行なうモデルを提案している。推定されたトピックは映像のセグメンテーションや後述する物体モデルの自動学習で利用する。トピック推定は、隠れ状態がトピックにあたり、言語、画像、音声情報の様々な素性が観測される隠れマルコフモデルで行なう。本研究では節を解析の基本単位とし、トピックは下ごしらえ、蒸す、ゆでる、揚げる、煮る、炒める、焼く、盛り付けの8つとする。基本素性は作業に関する発話を汎化した格フレームであり、それに加えてトピックが同一/異なることを示す、手がかり表現や語の連鎖などの談話素性を利用する。さらに、頑健の解析を実現するために、言語情報に加えて、映像情報としては背景画像、音声情報としては無音区間を利用する。実験を行なったところ、言語情報と映像情報の両方を利用することで、言語/映像の片方を利用するよりも精度向上がみられた。

第４章は「Automatic Object-Model Acquisition and Object Recognition」（物体モデルの自動学習と物体認識）と題し、料理映像で現われる食材の物体モデルを自動学習し、それを用いて物体認識を行なう手法を提案している。まず、物体がアップになっている画像を抽出し、その画像における注目領域を決定する。次に、発話タイプ認識、談話構造解析などの言語処理に基づき、画像の周辺の発話から重要な単語をキーワー

ドとして抽出し、注目領域と対応付ける。この際に、推定されたトピックが下ごしらえの部分から学習データを集めることにより、ノイズを軽減する。このような注目領域とキーワードを大量に収集することにより、物体モデルを構築する。物体モデルが構築された後、物体モデルの色情報と談話構造に基づく単語の重要度を考慮することにより、物体認識を行なう。2つの料理番組、計95時間分の映像から物体モデルを構築したところ、約100食材の物体モデルが構築でき、その精度は77.8%であった。また、そのモデルを利用して物体の認識を行なったところ、精度はF値で0.727であった。また、物体認識結果を省略解析の素性として取り入れたところ、言語だけで省略解析を行なうよりも精度が向上した。

第５章は「Video Retrieval System」（映像検索システム）と題し、実装した料理映像検索システムについて述べている。このシステムではユーザは自然言語で検索することができ、クエリと最も類似したテキストを含む映像を見ることができる。また、映像にアクセスしやすい手段をユーザに提供するために映像の要約を自動生成している。まず、トピック推定結果に基づき、映像のセグメンテーションを行ない、物体認識結果に基づき代表画像を抽出する。そして、発話タイプ認識・省略解析・談話構造解析・物体認識結果に基づき、重要な発話の抽出・整形を行ない、抽出した代表画像と重要発話を並べることで要約とする。要約生成の実験を行なったところ、映像の構造を十分捉えられていることが確認された。

第６章「Conclusions」（結論）では、本論文の主たる成果をまとめるとともに、今後の方向性について述べている。

以上を要するに、本論文は、作業教示映像の構造的理解を行なうための構造的言語処理とそれに基づく言語情報と映像情報の統合処理を提案し、大規模映像を利用した実験を通じてその有効性を示したものであり、電子情報学上貢献するところが少なくない。

よって本論文は博士（情報理工学）の学位請求論文として合格と認められる。