論 文 内 容 の 要 旨

博士論文題目　Linking Videos and Languages: Representations and Their Applications (映像とテキストを対応づける特徴量の開発およびその応用)

氏　名　大谷 まゆ

(論文内容の要旨)

映像理解はコンピュータビジョンにおける長年の目標であり、近年は自然言語を併用した研究が盛んになっている。自然言語を併用することで、映像の視覚的理解だけではなく、コンテンツベースでの映像検索や視覚障害者支援のための自動記述生成などの幅広い応用が考えられる。本論文では、映像と自然言語の双方を同一空間中で表現することにより、異なるモダリティを互いに関連付けた映像要約や映像検索の新しい手法を提案している。

第1章で導入、第2章で関連研究を述べた後、第3章で自然言語テキストを入力とした映像要約手法を提案している。具体的には、映像を特定のオブジェクトの有無により表現し、自然言語は前述のオブジェクトに対応する名詞の有無により表現することで、同一空間での表現を実現している。従来手法では要約映像中の内容を制御する手段を提供していないが、提案手法では自然言語テキストを入力としてユーザの意図を反映した要約映像の生成を実現している。映像と自然言語テキストが含まれるビデオブログ作成のシナリオの下で提案手法による要約映像の生成を客観的・主観的に評価することで、その有用性を検証している。

第4章では前章とは異なり、映像とテキストを共通の特徴空間にマッピングすることで、長時間映像の内容を可能な限り含めつつ短縮する映像要約手法を提案している。具体的には、映像と自然言語を同一空間に写像するニューラルネットワークを学習し、得られた写像を用いることで、短時間に分割された部分映像を高次元の空間中の点として表現する。これらの点をクラスタリングした上で、代表点に対応する部分映像を要約映像に含めている。種々の既存手法、および人手による要約映像と比較することにより、手法の有効性を示している。また、映像の検索などのタスクも合わせて深層学習による表現を評価している。

第5章では、自然言語クエリを用いた部分映像検索という新しい映像検索手法を提案している。具体的には深層学習を利用した映像の表現を拡張して映像の各フレームにおいて高次元のベクトルを出力することでこれを実現している。様々なニューラルネットワークの構成について検索の精度を客観的に評価する実験を実施し、手法の有効性を検証している。

第6章では本論文を総括し、今後の展望について述べている。

氏　名	大谷　まゆ

（論文審査結果の要旨）

本論文では、映像と自然言語の双方を同一空間中で表現することにより、異なるモダリティを互いに関連付けることを目的とし、映像の特徴量抽出や自然言語の構文解析に基づく古典的な手法と深層学習を利用した手法をそれぞれ提案している。古典的な手法では、映像を特定のオブジェクトの有無により表現し、自然言語は前述のオブジェクトに対応する名詞の有無により表現することで、同一空間での表現が可能である。深層学習を利用した手法は、映像と自然言語を同一空間に写像するニューラルネットワークを学習する。本論文では、それぞれの手法について、具体的な応用例を介してこれらの手法の有用性の検証・及び評価を実施している。本論文の主な成果は以下の 3 点である。

1. 長時間の映像を短くする映像要約には様々な手法が提案されているが、多くは生成された要約映像中に含まれる内容を直接制御する手段を提供していない一方、本論文では古典的な映像と自然言語の表現により、自然言語テキストを入力として要約映像の内容の制御を実現している。映像と自然言語テキストが含まれるビデオブログ作成のシナリオの下で提案手法による要約映像の生成を客観的・主観的に評価することで、その有用性を検証している。

2. 長時間の映像が持つ内容を可能な限り含めつつ短時間にまとめる映像要約は、前述のものとは目的を異にする映像要約の主要なタスクの一つである。本論文では、深層学習によって得られた写像を用いることで、短時間に分割された部分映像を高次元の空間中の点として表現し、得られた点をクラスタリングした上で、代表点に対応する部分映像を要約映像に含める手法を提案している。種々の既存手法、および人手による要約映像と比較することにより、手法の有効性を示している。また、映像の検索などのタスクも合わせて深層学習による表現を評価している。

3. 映像検索では、多くの場合短時間の映像を対象としていたが、現実には映像の一部分の検索が要求される場合もある。本論文では、自然言語クエリを用いた部分映像検索という新しい映像検索の問題を提唱し、深層学習を利用した映像の表現を拡張して映像の各フレームにおいて高次元のベクトルを出力することで、これを実現している。様々なニューラルネットワークの構成について検索の精度を客観的に評価する実験を実施し、手法の有効性を検証している。

以上のとおり、映像の意味解析というパターン認識の分野における主要な問題に対して、自然言語処理の援用によるアプローチで取り組んでおり、実験によってそれぞれの応用例においてその有用性・有効性を示している点で、当該分野に対する貢献を認めることができる。本論文の内容は、英文論文誌と国際会議などで発表されている。

よって、本論文は博士（工学）の学位論文として価値のあるものであると認める。