（ 続紙 1 ）

京都大学	博士（情報学）	氏名	西村太一
論文題目	Procedural Text Generation from Instructional Videos （作業映像からの手順書生成）		

（論文内容の要旨）

本論文は、マルチメディアアーカイブの構築のための映像と言語の統合処理技術の開発を目標に、作業映像からの手順書生成を行うアプローチについての研究結果をまとめたものであり、全6章から構成されている。

第1章は序論であり、本論文の背景と目的を概観している。映像と言語の対応関係を獲得する研究はマルチメディア分野で長年取り組まれてきた課題であるが、既存研究では手順書、音声書き起こし、映像の3つが入力として必要であった。深層学習の登場により、映像を直接言語化するVideo captioningが発展した。しかしながら、こうした技術は数十秒程度の短い映像に焦点を当てており、時間的に長く、多種多様な物体や動作が現れる作業映像に直接適用することはできない。本論文では、Video captioning技術を作業映像から手順書を生成することに拡張することを目指す。映像から手順書を生成する過程を(1)映像から作業を達成する上で重要なシーン(イベントと呼ぶ)を抽出し、(2)イベントに対して文を生成する2つの課題へと定式化し、その上で3つの課題及びそのアプローチについて述べている。

第2章では、映像と言語の対応関係を獲得することを目指した過去のマルチメディア研究、そして深層学習の登場により急速に発展している視覚と言語の統合処理研究について俯瞰し、本論文の位置付けを行っている。マルチメディア研究の点からは、過去の映像と言語の対応関係を獲得する研究と比較し、本論文が映像を直接言語化する点における新規性を述べている。また、視覚と言語の統合処理研究の点からはVideo captioning技術と比較して、時間的に長く、多種多様な物体や動作が現れる作業映像への適用が挑戦的でかつ新規であることを論じている。

第3章では、イベント列と映像で使われている材料を入力として手順書を生成する課題に取り組む。従来の映像キャプショニングモデルを本問題に適用すると、正しくない材料を生成したり、過度に一般化された文を生成したりすることが分かった。この原因として、本論文では、作業映像では人の操作によって材料の性質が変化し、結果見た目の変化を伴うためであると考えた。この仮定を検証するべく、手順書理解の研究で開発された、材料の状態変化をベクトル空間上で表現する手法をマルチモーダルに拡張し、手順書生成モデルへ組み込んだ。その結果、従来のVideo captioning手法と比較して正しく材料や動作を含んだ手順書を生成できることを確認した。

第4章では、作業映像からイベント列の抽出と手順書生成を同時に学習する問題を取り扱う。入出力が共通しているという点から、この課題はコンピュータビジョン領域で取り組まれてきたDense Video Captioning (DVC)と類似している。DVCは映像から不足なくイベントを抽出することに焦点を当てている一方で、本研究で取り組む課題は映像から必要な数、正しい順番で予測し文を生成することを重視している。本研究では、これをストーリー性と定義し、ストーリー性を考慮した手順書生成に取り組む。DVCの出力を分析した結果、出力のイベント集合には正解イベントと類似するものは存在するが、生成された文は正しいものではないことが分かった。そのため、

本論文ではDVCの出力するイベント集合をイベント候補とみなし、そこから再帰的にイベントを選択しながら文を生成する手法を提案している。また、第3章の知見を活かし、材料を追加で入力を与える拡張モデルも提案している。実験の結果、提案モデルはDVCの既存手法と比べてストーリー性に沿った手順書生成を行えること、及び拡張モデルはより正しく手順書を生成できることを確認している。

第5章では、今まで手順書生成の対象分野を料理や裁縫といった日常的なものから言語化することの重要性の高い分野へ拡張する課題に取り組む。本研究では、再現性の危機の観点から言語化することの需要の高い生化学分野へ拡張することを目指す。実験映像を言語化したり、言語から実験映像を検索したりできるような魅力的な応用が考えられる一方、生化学分野では映像と言語の利用可能なデータセットが存在しない。この問題を解決するために、一人称視点での映像と言語のデータからなるBioVL2データセットを構築した。このデータセットには、一人称の実験映像に対して、(1)プロトコル(実験手順書)の各手順と映像中のイベントの対応関係と(2)映像中の物体矩形の2種類のアノテーションを付与している。構築したデータセットをもとに、応用課題として映像からプロトコルを生成する課題に取り組み、現時点での到達点と今後の生化学を対象としたマルチメディア研究の方向性について議論している。

第6章は結論であり、本論文で得られた成果を要約している。即ち本論文は、マルチメディアアーカイブの構築のための映像と言語の統合的処理の実現を目標に、作業映像から手順書を生成する課題に取り組み、その有効性を実験的に示したものである。本論文では最後に、実用上での本研究の限界点について論じ、今後の展望として一人称映像からの手順書生成、分野に非依存な手法の開発、そして学習して得た材料ベクトルの他課題への転移を挙げ、本論文を結んでいる。

（続紙 ２ ）

（論文審査の結果の要旨）

本論文は、マルチメディアアーカイブの構築のための映像と言語の統合処理技術の開発を目標に、作業映像からの手順書生成を行うアプローチについての研究結果をまとめたものであり、得られた主な成果は次の通りである。

１． 手順書に対して材料の状態変化をベクトル空間上で表現する手法をマルチモーダルに拡張し、物体の見た目の変化を伴う作業映像に対して、captioningを行う首相を提案・実装し、従来のVideo captioning手法と比較して正しく材料や動作を含んだ手順書を生成できることを確認した。

２．前成果を活用し、作業映像からイベント列の抽出と手順書生成を同時に学習する方法を提案・実装した。提案モデルはDense Video Captioning (DVC)の既存手法と比べてストーリー性に沿った手順書生成を行えること、及び拡張モデルはより正しく手順書を生成できることを確認した。

３．ここまでの成果を生化学分野へ拡張することを試みた。このために、まず一人称視点での映像と言語のデータからなるBioVL2データセットを構築した。構築したデータセットをもとに、実験映像からプロトコルを生成する課題に取り組み、現時点での到達点と今後の生化学を対象としたマルチメディア研究の方向性について議論した。

以上、本論文は、調理や化学実験の実施映像とから手順書を生成するという課題に対して適切な解決方法を提案し、既存の手法を超えることを実験的に確認している。加えて、新たな対象（生化学実験）に対してデータセットを構築・公開したものであり、学術上・実応用上寄与するところが少なくない。よって、本論文は博士（情報学）の学位論文として価値あるものと認める。また、令和５年８月２日に実施した論文内容とそれに関連した口頭試問の結果、合格と認めた。なお、インターネットでの全文公表を行うことについて支障がないことを確認した。