様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　　元 年　　５ 月 ２３ 日現在

機関番号：１４４０１
研究種目：若手研究(B)
研究期間：2016～2018
課題番号：１６Ｋ１６０８６

研究課題名（和文）大量の映像群からテキストの内容に沿った映像を生成する映像要約手法の開発

研究課題名（英文）Video summarization from a set of videos controlled by text input

研究代表者
中島　悠太（Nakashima, Yuta）

大阪大学・データビリティフロンティア機構・准教授

研究者番号：７０６３３５５１

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：映像編集の自動化を目的として、映像からその要約映像生成する技術を構築することを目標とする。映像には様々なシーンにおける様々な粒度のイベントが含まれており、どのイベントを要約映像に含めるかは、映像要約の応用先によって異なる。本研究では、ユーザからの入力などによって映像に含まれるフレームの重要度を決定するものと、意味内容をなるべくカバーしつつ冗長性を低減させるものの二つのアプローチを考え、それぞれの手法を提案し、その有用性について実験的に示した。また、映像要約の評価手法についても検討を進め、新たな評価手法を提案した。

研究成果の学術的意義や社会的意義
近年、スマートフォンやデジタルカメラなどで日々大量の映像が撮影されている。本研究は、このような映像に対して短時間でその内容を閲覧可能な映像要約手法を提案しており、映像閲覧時のユーザの負荷の軽減や送受信される映像サイズの削減などの点で高い有用性を持つと考える。加えて、特に映像要約の評価手法については、広く用いられるデータセットで利用される評価手法の問題点を明確にしているという点において、今後の映像要約の研究に大きな影響を与えるものであり、学術的意義も大きいと考える。

研究成果の概要（英文）：This work aims at automating video editing and establishes techniques for automatically generating a video summary from an original video or a set of them. Generally, a video contains various events occurring in various scenes, and it is not obvious that which events should be included in the resulting video summary. In this work, we considered two approaches for video summarization: One approach determines each frame's importance based on certain types of user input. The other approach attempts to reduce the redundancy in the resulting video summary while covering the content of original video as much as possible. For respective approaches, we proposed video summarization methods and experimentally demonstrated their effectiveness. We also reconsider the evaluation of video summarization method and developed a new method for evaluation.

研究分野：パターン認識、コンピュータビジョン

キーワード：映像要約　深層学習　畳み込みニューラルネットワーク　重要領域推定　要約映像の評価

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

1.　研究開始当初の背景

近年、スマートフォンやデジタルカメラなどで大量の映像が撮影されている。これらの映像を活用する方法としてビデオブログが考えられる。ビデオブログは vlog とも呼ばれており、映像とその映像に関するテキストで構成される。 ビデオブログの生成のためには、まず撮影された映像の内容を確認した上でビデオブログの内容に沿うように映像を編集するとともに、ブログ用のテキストを記述する。このプロセスにおいて、テキストの記述は容易である一方で、映像編集には多大な手間を要する。このため、大量の映像を簡便にビデオブログとして活用するためには、特に映像編集の自動化が必要である。

２．研究の目的

本研究では、映像編集の自動化を目的として、映像からその要約映像生成する技術を構築することを目標とする。映像には様々なシーンにおける様々な粒度のイベントが含まれており、どのイベントを要約映像に含めるかは、映像要約の応用先によって異なるが、大きく二つのアプローチが考えられる。第 1 のアプローチは、ユーザの嗜好などを基に要約映像に含まれる映像を制御するもの、第 2 のアプローチは、要約映像の冗長性を小さくしつつ映像の意味内容を広くカバーしようとするものである。本研究では、第 1 のアプローチをメインとして、ユーザが自然言語テキストを入力することにより、要約映像を生成する。具体的には (1) 自然言語テキスト-映像間の類似度計算手法の確立、(2) テキストに基づく要約映像の自動生成手法の構築の要素について検討した。加えて、第 2 のアプローチについても、(3) 自然言語テキスト-映像間の類似度計算に基づく映像要約手法を提案した。要約映像の評価については、(4) 既存のデータセットで用いられる評価方法に関して検討した。さらに、第 1 のアプローチに関連して、(5) スポーツ映像の要約手法を開発した。

３．研究の方法

(1) 自然言語テキスト-映像間の類似度計算手法の確立: ユーザが自然言語テキストを利用して要約映像に含まれる映像を制御する際には、映像とテキストの類似度を算出する必要がある。本研究では、ディープニューラルネットワーク（Deep Neural Network; DNN）を用いて、テキストと映像の共通空間への写像を学習する手法を提案した。この手法では、再帰型ニューラルネットワーク（Recurrent Neural Network; RNN）を用いてテキストを写像する際に、その詳細が失われるという問題を解消するために、テキストをクエリとしてウェブ画像検索することにより得られる画像から得られる特徴量をテキストと合わせて写像する。対応する映像とテキスト（ウェブ画像）が共通空間中のなるべく近い点に、対応しないものは離れた点に写像されるように DNN を学習し（図 1)、共通空間中での距離により類似度を定義する（[学会発表②]）。

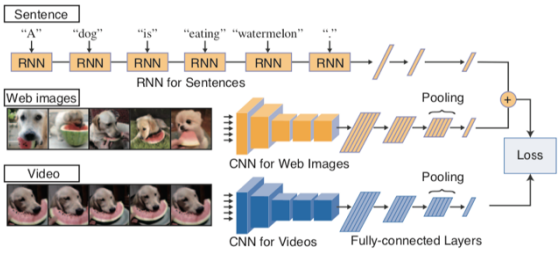

図 1: 自然言語テキスト-映像間の類似度計算手法

(2) テキストに基づく要約映像の自動生成手法の構築: ユーザが入力するテキストに従って内容を制御可能な映像要約手法として、ここでは図 2 に示すように、テキストに含まれる名詞集合と、映像に登場する物体を検出することで得られる物体ラベル集合の一致度によりテキストと映像の間の類似度を定義し、この類似度を最大化するように部分映像を選択する最適化問題として要約映像の自動生成を定式化した（[雑誌論文②]）。

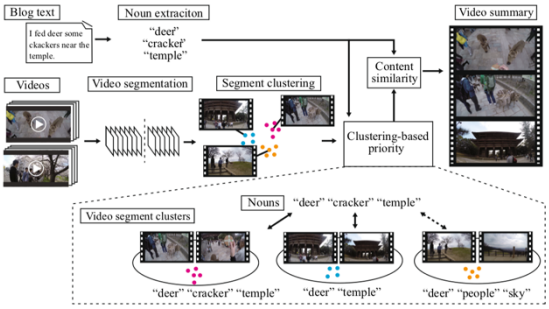

図 2: テキストに基づく映像要約の自動生成手法

(3) 自然言語テキスト-映像間の類似度計算に基づく映像要約手法: 第 2 のアプローチとして、(1) で確立したテキスト-映像間の類似度計算手法を用いた手法を開発した。この手法では、部分映像を (1) と類似の方法で訓練したネットワークに入力して得られるベクトルは、その部分映像の意味内容を表現する特徴量であると考え、図 3 のように部分映像をクラスタリングし、各クラスタを代表

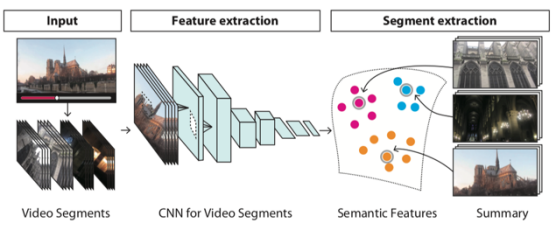

図 3: 自然言語テキスト-映像間の類似度計算に基づく映像要約手法

する部分映像を要約映像に含める（[学会発表③④]）。

（4）既存のデータセットで用いられる評価方法に関する検討：一般に映像要約は、その目的などによって様々な評価基準を考えることができる。第2のアプローチについては、多くの研究で、映像を部分映像に分割した上で、各フレームに対して重要度を算出し、その値に基づいて要約映像を生成する。評価には公開されたデータセットが広く利用されており、再現率と適合率を基にしたFスコアが用いられる。この課題では、図4のようにランダムな部分映像への分割と重要度の値によって得られた要約映像の評価値が最新の手法に近いものとなるという既存のデータセットを利用した評価の問題点を明確化し、映像要約手法の評価を可視化することにより、評価結果の解釈を容易にする手法に加えて、新たな評価基準を提案した（[学会発表⑨]）。

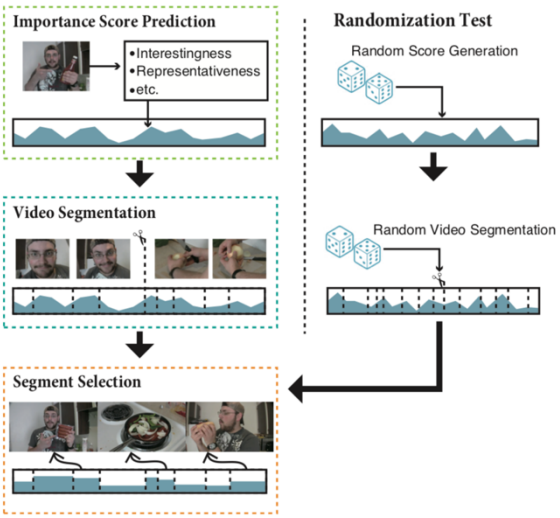

図 4: ランダムな部分映像分割と重要度による要約映像生成

（5）スポーツ映像の要約手法の開発：映像要約の応用の一つとして、要約映像内に含める部分映像の基準が比較的明確であると考えられるスポーツ映像の要約手法を開発した（[雑誌論文①③][学会発表①]）。この手法では、剣道の試合を撮影した映像を対象とし、選手の姿勢検出に基づく動作認識によって部分映像を剣道の動作クラスへと分類する識別器をDNNで構築した上で、この識別器の中間層の出力を特徴量としてユーザが指定した重要な部分映像をフレーム単位で見つけるように識別器を学習した。要約映像は、この識別器の出力に従って生成される（図5）。

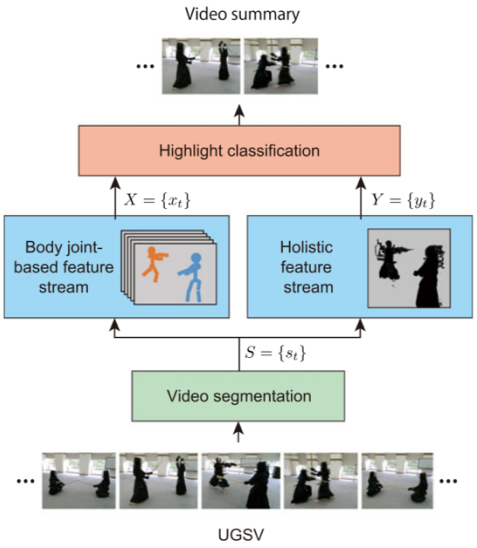

図 5: スポーツ映像の要約手法

上記（1）～（5）の課題に加えて、（2）への適用を考え、テキストを利用した部分映像の検索手法（[学会発表⑤⑥⑦]）や、映像中の重要領域を検出する手法（[雑誌論文④]）を提案した。

４．研究成果

（1）自然言語テキスト-映像間の類似度計算手法の確立：YouTube 映像に関するデータセットを用いて提案するDNNに基づく類似度計算手法を評価した。このデータセットは、1970件の映像に対して合計で8万件の英語の説明文が付与されている。説明文から映像を検索するタスクでは、検索結果のトップ10件以内に正解の映像が含まれる割合を表すR@10の評価指標で、提案手法の一つが35%（チャンスレート：1.5%、比較手法20%）を達成した。映像から説明文を検索するタスクでは、同じ評価指標で38%（チャンスレート1.3%、比較手法17%）となった。この結果から、提案手法は高い性能を持つことがわかる。

（2）テキストに基づく要約映像の自動生成手法の構築：この課題では、テキストに基づく映像要約という新しいタスクを提案しており、確立された評価指標は存在しない。ここでは、与えられた3つのテキストから自動的に生成された要約映像が、それぞれテキストに対して適切かを1から5点のアンケートにより調査した。ベースラインとして、元の映像から部分映像をランダムに抽出したもの、映像をクラスタリングし、それぞれのクラスタの代表点を要約映像に含めるものを利用した。結果、テキストの内容によってばらつきはあるものの、提案手法は高い評価を得た。

（3）自然言語テキスト-映像間の類似度計算に基づく映像要約手法：25本の映像を含むデータセットにより要約映像を評価した。このデータセットは、1本の映像について少なくとも15本の手作業で作成された要約映像が提供されており、生成された要約映像は手作業で作成された要約映像とどの程度一致するかを示すFスコアにより評価される。結果、提案手法は、教師なしの手法の中で最も高いスコア（0.183）となった。手作業で生成された要約映像に基づいて映像要約のためのモデル学習する教師ありの手法は、Fスコアが0.234、手作業で生成された要約

映像を Leave-One-Out Cross Validation で評価したところ、最大値が 0.409、最小値が 0.179、平均値が 0.311 となった。教師なしの手法であるにも関わらず、手作業で生成された要約映像の評価値の最小値よりも優れた性能が得られた。

(4) 既存のデータセットで用いられる評価方法に関する検討: 本課題では、要約映像の生成時に算出されるフレームごとの重要度から、手作業で作成された要約映像群が与えられた時に、自動生成された要約映像がどの程度優れたものかを、グラフによって表現する可視化手法を提案した (図 6)。これに加えて、重要度に基づいて要約映像を生成する手法については、生成された要約映像自体ではなく、その中間表現である重要度に基づく評価が必要であると考え、重要度の順位相関係数を利用した評価手法を提案した。

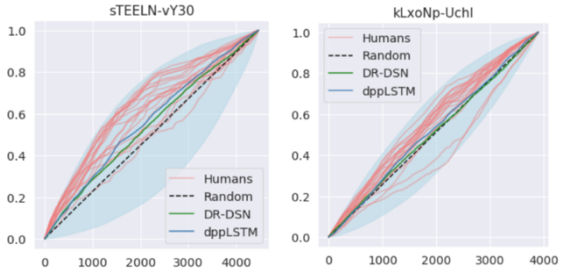

図 6: 映像要約における重要度の可視化

(5) スポーツ映像の要約手法の開発: 剣道の試合を Kinect v2 を利用して撮影した合計 246 分の長さの映像に対して、剣道経験者と未経験者それぞれ複数名が映像中の重要箇所に対してラベルを付与し、このラベルを使って DNN の訓練、および評価を実施した。提案する選手の 3 次元の姿勢に関する特徴量と、映像全体に関する特徴量の双方を統合したネットワークを利用した手法は、剣道経験者と未経験者のラベルについてそれぞれ F スコアで 0.58、0.85 となった。K-mean クラスタリングによる手法 (それぞれ 0.28、0.61)、ガウス混合分布を用いた隠れマルコフモデルを利用した手法 (それぞれ 0.44、0.79) に比べて高い値となり、提案手法の有効性が示せたと考える。

５．主な発表論文等

〔雑誌論文〕(計 4 件)

① Mayu Otani, Atsushi Nishida, Yuta Nakashima, Tomokazu Sato, and Naokazu Yokoya, "Finding important people in a video using deep neural networks with conditional random fields," IEICE Transactions on Information and Systems, vol. E101.D, no. 10, pp. 2509-2517, 2018 (査読有).
DOI: 10.1587/transinf.2018EDP7029

② Antonio Tejero-de-Pablos, Yuta Nakashima, Tomokazu Sato, Naokazu Yokoya, Marko Linna, and Esa Rahtu, Summarization of user-generated sports video by using deep action recognition features," IEEE Transactions on Multimedia, vol. 20, no. 8, pp. 2000-2011, 2018 (査読有).
DOI: 10.1109/TMM.2018.2794265

③ Mayu Otani, Yuta Nakashima, Tomokazu Sato, and Naokazu Yokoya, "Video summarization using textual descriptions for authoring video blogs," Multimedia Tools and Applications, vol. 76, no. 9, pp. 12097-12115, 2017 (査読有).
DOI: 10.1007/s11042-016-4061-3

④ Antonio Tejero-de-Pablos, Yuta Nakashima, Naokazu Yokoya, Francisco-Javier Díaz-Pernas, and Mario Martínez-Zarzuela, "Flexible human action recognition in depth video sequences using masked joint trajectories," EURASIP Journal on Image and Video Processing, vol. 2016, 20 pages, 2016 (査読有).
DOI: 10.1186/s13640-016-0120-y

〔学会発表〕(計 9 件)

① Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkilä, "Rethinking the evaluation of video summaries," Proc. Computer Vision and Pattern Recognition (CVPR) 2019.

② Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Naokazu Yokoya, "Linking videos and languages: Representations and their applications," 情報処理学会 研究報告, CVIM-212-38, 2018.

③ Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkilä, "Finding video parts with natural language," 情報処理学会 研究報告, CVIM-211-7, 2018.

④ Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkilä, "Fine-grained video retrieval for multi-clip video. Proc. Workshop on Closing the Loop Between Vision and Language (CLVL) at ICCV, 2017.

⑤ Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkilä, "Video question answering to find a desired video segment," Proc. Open Knowledge Base and Question Answering

Workshop (OKBQA) at SIGIR, 2017.
⑥ Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Naokazu Yokoya, “Unsupervised video summarization using deep video features,” 画像の認識・理解シンポジウム (MIRU), 2017.
⑦ Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, Naokazu Yokoya. Video summarization using deep semantic features. Proc. 13th Asian Conference on Computer Vision (ACCV), 2016.
⑧ Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, Naokazu Yokoya, “Learning joint representations of videos and sentences with web image search,” Proc. 4th Workshop on Web-scale Vision and Social Media (VSM) at ECCV, 2016.
⑨ Antonio Tejero-de-Pablos, Yuta Nakashima, Tomokazu Sato, Naokazu Yokoya, “Human action recognition-based video summarization for RGB-D personal sports video,” Proc. 2016 IEEE International Conference on Multimedia and Expo (ICME), 6 pages, 2016.

〔その他〕

ホームページ等: https://www.n-yuta.jp/project/video-summarization/