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　４ 年　 ６ 月 ２８ 日現在

機関番号：３２６８５
研究種目：基盤研究(C)（一般）
研究期間：2018～2021
課題番号：１８Ｋ１１３６２

研究課題名（和文）未知の概念を含むクエリ文を用いた大規模映像からの詳細映像検索

研究課題名（英文）Fine-grained video retrieval from large-scale video using query sentences containing unknown concepts

研究代表者
植木　一也（Ueki, Kazuya）

明星大学・情報学部・准教授

研究者番号：８０５８０６３８

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：インターネット上に日々アップロードされる多種多様な映像中から，新しく生まれるトレンドや新しい手口の犯罪等，新規の概念を含む未知のクエリ文に合致する映像を検索する技術に取り組んだ．大量の説明文付きの画像を用いて画像と言語の埋め込みを行い，学習されたモデルを動画フレームに対して適用する手法を検討した．作成したシステムを，米国国立標準技術研究所主催の国際競争型映像検索・評価ベンチマーク（TRECVID）に提出し，大規模映像を用いて評価を行ったところ，幅広いクエリ文に対して高精度で映像を検索できることが確認できた．

研究成果の学術的意義や社会的意義
ラベル付きの画像・映像データの整備と，ディープラーニング技術の進展に伴い，ある特定の物体・シーン・動作等のキーワードに合致した画像や映像の検索が実現されつつある．しかしながら，ライフスタイルの変化により，複数かつ新しい概念を含んだクエリ文を用いた詳細映像検索の実現が期待されている．本研究成果により，幅広く複雑なクエリ文に対して高精度に映像を検索することが可能であることから，新しく生まれる未知の概念が含まれている場合においても，その説明文をクエリとして入力することで，必要となる映像を即座に検索できることが期待される．

研究成果の概要（英文）：We worked on a technique for retrieving videos that match unknown query sentences containing new concepts, such as newly emerging trends and new methods of crime, from the wide variety of videos uploaded to the Internet every day. We investigated methods to apply the trained models to video frames by embedding images and language using a large number of images with captions. We evaluated our method on a large scale of videos in the international competitive video retrieval and evaluation benchmark (TRECVID) organized by the National Institute of Standards and Technology (NIST), and confirmed that our method could retrieve videos with high accuracy for a wide range of query sentences.

研究分野：知覚情報処理

キーワード：映像検索　クエリ文　TRECVID　未知の概念　画像/言語の同時埋め込み

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

大量の正解付き画像・映像の整備と，ディープラーニング技術の進展により，画像中の一般物体やシーンの認識，動画中の人の動作やイベントの検出のタスクにおいて，人間の能力を上回る精度を達成したといわれている．実際，2017 年まで毎年実施された ImageNet 画像認識コンペティションでは，人の認識率よりも高いという報告があった．申請者が毎年参加している国際競争型大規模映像検索・評価 TRECVID ベンチマークにおいても，年々飛躍的に認識精度が向上している．これらは，物体・シーン・動作等の概念が事前に与えられ，その概念に対する正解付きの学習データを大量に用いることで実現している．しかしながら，日々新しい概念が生まれている社会環境において，単純に既知の概念のみを扱うだけでは本当に獲得したい情報を得ることが難しくなってきている．そのため，特定の物体／人の行動／シーン等の複数の概念を同時に扱うことができ，さらに学習データ中にない新しい概念についても認識可能とすることは取り組む価値があると考えた．

２．研究の目的

画像認識の分野において，ある概念を含む学習画像が存在しない場合でも，その概念に付与されている属性情報を活用することで未知の概念を認識するゼロショット学習の技術が存在する．これに対して本研究提案の手法では，検出したい概念に対し，その度ごとに人間が知識を与えるのではなく，大量の画像・映像・言語資源を元に自動で獲得するという特徴をもつ．以上の研究の過程で，画像・映像と言語の特徴空間に相互に自由に変換するための新たな技術基盤ができ，画像・映像と言語のより深い関係性の獲得や意味理解ができるものと期待している．

３．研究の方法

正解付きの学習データがない環境下での映像検索を実現するため，言語・画像・映像資源を用いて，以下の４つの研究項目を実施する．

- ［研究項目１］少量の正解付きデータが得られるという条件での検索技術の検討

クエリ文に対応する正解付きの画像・映像サンプルは，多少であれば入手できる場合もある．その場合，ディープラーニングのように大量の学習サンプルを使った学習手法は用いることはできないが，学習済みの認識器から算出される各概念のスコアを用いて，どういった概念を含んでいる可能性があるかを知ることができる．これにより，関連すると思われる検索キーワードの候補をユーザに提示することや，完全に自動で関連する概念を抽出することができるようになる．そのため，実際に少量の学習サンプルが利用できる仮定のもとでの実験により，関連する概念をどの程度正確に抽出できるかを確認するとともに，誤って抽出された概念に対しては，その原因を明らかにしていく．

- ［研究項目２］言語・画像・映像資源を用いて新たな検索キーワードの候補抽出を行う手法の確立

クエリ文を単語に分解して解析する手法を確認するため，大規模の言語・画像・映像資源を用いることで，クエリ文に含まれる概念を自動抽出する仕組みを構築する．言語資源としては，整備された大規模言語コーパスや，インターネット上に存在する言語資源を活用する．文の中で使われ方が類似している単語を獲得する方法，文章中に共起しやすい単語を抽出する方法，単語の概念辞書（WordNet）を利用して同義語や単語の定義を活用する方法により，クエリ文中にある未知語に対応する概念を補完する．画像・映像の特徴抽出については，学習済みのディープラーニングの認識モデルを用いることで，複数の概念の共起関係と新たな概念を獲得できるかどうかを検証する．

- ［研究項目３］クエリ文から直接的に映像を検索する手法の検討

クエリ文を単語に分解して解析するだけでなく，クエリ文を直接扱い，映像を検索する手法を検討する．映像を入力したときに出力される説明文や学習済みのニューラルネットワークの中間表現を用いて特徴の類似度を計算する手法を検討する．ここでは単純に文の意味の類似性を見るのではなく，画像や映像を表現する文同士の類似性を知る必要がある．以上のことから，映像と説明文との間の関係性を理解し，未知の概念を含むクエリ文であっても直接的に映像を検索できる機能を実現する．

- ［研究項目４］全体の統合システムを作成して評価を実施

上記で開発した手法を統合し，映像検索システムを作成して評価を行う．システム評価は，

TRECVID 等の 100 万を超える映像を含んだ大規模データベースを用いて定量的な評価を行い，手法の有効性を確認する．

４．研究成果

[研究項目１]については，クエリ文に完全に合致した少量のデータ（画像・映像）を収集し，そのデータがすでにある概念とどの程度相関があるかを分析することにより，クエリ文中に直接に記述されていない潜在的な概念を新たに獲得できることを確認した．例えば，クエリ文「destroyed buildings」に合致する画像を準備し，各画像を概念識別器で評価して，関連のある概念の抽出を試みた．その結果，「destroyed buildings」に近い意味の「ruin」が選択されていることがわかった．また，相関の高い概念をさらに確認してみたところ，「garbage heap」「dump」等，直接的に関係ないが，画像の特徴としては近い（視覚的に近い）概念が選ばれていることがわかった．これらの概念を利用して映像検索をすることにより，図１に示すように映像検索の精度を向上できることが明らかになった．

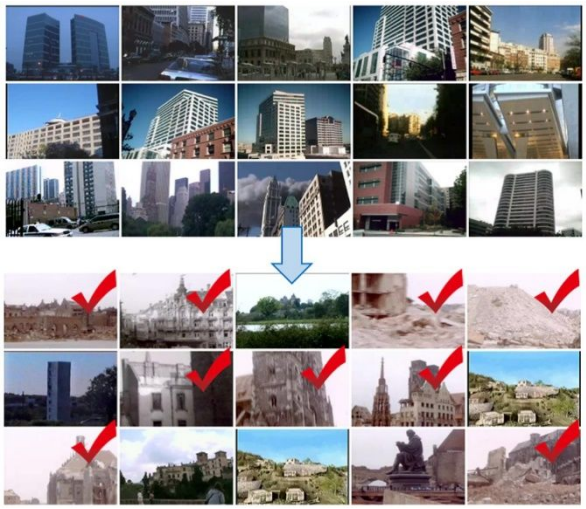

図１　「destroyed building」の映像検索結果．チェックマーク付きが正解の映像．

[研究項目２]については，Google 社が開発した Universal Sentence Encoder [1] を活用し，クエリ文に近い概念を直接的に取得することを試みた．具体的には，クエリ文をチャンクに分けたフレーズまたはクエリ文全体をエンコードしたベクトルと，概念識別器名との対応付けを行うことにより，新たな潜在的概念の獲得を目指した．また，概念識別器名は，英語の概念辞書（意味辞書）である WordNet と紐付いていることから，単語の上位概念や辞書の定義文も活用することで潜在的概念を獲得した．この手法により，検索キーワードと識別器名の対応付けを行う際，同じ識別器名をもつが意味が違った概念識別器を誤って選んでしまうケースを軽減できることがわかった．

[研究項目３]は，[研究項目２]において，大規模な言語・画像・映像データベース，特にその中でも言語資源の単語数の豊富さが映像検索の精度を左右することがわかったため，画像に対応する説明文（キャプション）が付与されているデータベースである MS COCO [2]（画像数：約 120,000，キャプション数：約 600,000），flickr 8k [3]（画像数：約 8,000，キャプション数：約 40,000），flickr 30k [4]（画像数：約 30,000，キャプション数：150,000），Conceptual Captions [5]（画像数：約 3,000,000，キャプション数：約 3,000,000）を収集して大規模データベースを構築した．次にニューラルネットワークを用いて画像と説明文を同一空間上に写像する画像/言語の同時埋め込みモデルの学習を実施した．TRECVID ベンチマークの大規模映像と実際にベンチマークで出題されたクエリ文を用いて，この手法の有効性を確認した．画像/言語の同時埋め込み手法では，大量のキャプション付きの画像を学習してモデルを作成したことから，事前に学習された概念識別器を活用する手法（コンセプトベースの手法）でカバーできない概念の多くを補完できることがわかった．特に図 2 に示すような「a crowd of people」のようなフレーズに対応できることが確認できた．

図 2　「a crowd of people」の映像検索結果．

また，近年提案されている画像/言語の同時埋め込み手法である Improving Visual-Semantic Embeddings (VSE++) [6], Graph Structured Matching Network (GSMN) [7], Object Semantics Aligned Pre-training (Oscar) [8], Contrastive Language-Image Pre-Training (CLIP) [9]を用いて，映像検索への応用を検討した．VSE++，GSMN については，説明文が付与されている画像約3,000,000枚を画像/言語の埋め込みの学習に利用した．Oscar と CLIP

については，自前で収集した画像よりも多くの画像を学習したモデルが公開されているため，それらを利用して映像検索に活用する手法についても検討を行った．

[研究項目４]のシステム全体の評価は，上記の研究項目で得られた知見をもとに，作成したシステムを TRECVID 映像検索ベンチマークに提出することにより行った．事前に学習された概念識別器を活用するコンセプトベースの手法と，画像と説明文を同一空間上に写像する手法を統合したシステムを評価したところ，検索精度が大幅に向上するなど，個々の手法の相補性が確認できた．また，GSMN で利用されている物体検出ベースの特徴抽出方法は，クエリ文からの映像検索においても有効であることがわかった．CLIP のように大量の説明文付きの画像を学習したモデルは汎用性が高く，どのようなクエリ文が入力された場合においてもロバストに検索が可能であった．説明文付きの映像データセットは，説明文付きの画像データセットに比べて極端にデータ数が少ないことから，現時点では，映像からフレーム画像を抽出して画像/言語の埋め込みモデルを用いる手法の方が，映像を高精度に検索できるということが明確となった．最終的には，複数の画像/言語同時埋め込みモデルによる手法を統合することにより，TRECVID ベンチマークにおいて世界一位のシステムとほぼ同等の精度の映像検索を達成し，新しく作成されるクエリ文に対しても頑健に映像検索できることを確認した．

＜引用文献＞

[1] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. GuajardoCespedes, S. Yuan, C. Tar, Y.-H. Sung, B. Strope, and R. Kurzweil, “Universal Sentence Encoder,” arXiv:1803.11175, 2018.

[2] T.-Y. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: Common Objects in Context,” arXiv:1405.0312, 2014.

[3] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, “Collecting Image Annotations Using Amazon's Mechanical Turk,” Proc. of the NAACLHLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk, pp.139-147, 2010.

[4] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions,” Transactions of the Association for Computational Linguistics. vol.2, pp.67-78, 2014.

[5] P. Sharma, N. Ding, S. Goodman, and R. Soricut, “Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning,” Proc. of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 2556-2565, 2018.

[6] F. Faghri, D. J. Fleet, R. Kiros, and S. Fidler, “VSE++: Improved Visual-Semantic Embeddings,” arXiv:1707.05612, 2017.

[7] C. Liu, Z. Mao, T. Zhang, H. Xie, B. Wang, Y. Zhang, “Graph Structured Network for Image-Text Matching," In Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[8] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, J. Gao, “Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks," In Proc. of European Conference on Computer Vision (ECCV), 2020.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, “Learning Transferable Visual Models From Natural Language Supervision," arXiv:2103.00020, 2021.

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 植木 一也，平川 幸司，菊池 康太郎，小林 哲則	４．巻 84
２．論文標題 複雑のコンセプトを含むクエリ文からのゼロショット映像検索 - TRECVID AVS タスクにおける成果と課題 -	５．発行年 2018年
３．雑誌名 精密工学会誌	６．最初と最後の頁 983-990
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計14件（うち招待講演　0件／うち国際学会　3件）

１．発表者名 セイエドネシャド ロスタム，武藤 良，植木 一也，堀 隆之，金 容範，鈴木 裕真
２．発表標題 服装の色を用いた人物検索に向けた学習済みモデルの活用
３．学会等名 動的画像処理実用化ワークショップ（DIA2021）
４．発表年 2021年

１．発表者名 山本 啓斗，武藤 良，植木 一也，堀 隆之，金 容範，鈴木 裕真
２．発表標題 少数画像をもとにした顔属性データセットの拡張
３．学会等名 動的画像処理実用化ワークショップ（DIA2021）
４．発表年 2021年

１．発表者名 Kazuya Ueki, Ryo Mutou, Takayuki Hori, Yongbeom Kim, Yuma Suzuki
２．発表標題 Waseda_Meisei_SoftBank at TRECVID 2020: Ad-hoc Video Search
３．学会等名 Notebook paper of the TRECVID 2020 Workshop
４．発表年 2020年

１．発表者名
武藤 良，セイエドネシャド ロスタム，植木 一也，堀 隆之，金 容範，鈴木 裕真

２．発表標題
学習済みモデルを用いた大規模映像データにおける特定の色の着衣をつけた人物の検索

３．学会等名
ビジョン技術の実利用ワークショップ(ViEW2020)

４．発表年
2020年

１．発表者名
セイエドネシャド ロスタム，武藤 良，植木 一也

２．発表標題
OpenPoseを用いた特定の色の服を着た人物の検出

３．学会等名
第26回画像センシングシンポジウム(SSII2020)

４．発表年
2020年

１．発表者名
Kazuya Ueki, Takayuki Hori

２．発表標題
Comparison and Evaluation of Video Retrieval Approaches Using Query Sentences

３．学会等名
the 3rd International Conference on Machine Vision and Applications (ICMVA 2020)（国際学会）

４．発表年
2020年

１．発表者名
Kazuya Ueki, Takayuki Hori, and Tetsunori Kobayashi

２．発表標題
Waseda_Meisei_Meisei at TRECVID 2019: Ad-hoc Video Search

３．学会等名
the TRECVID 2019 Workshop（国際学会）

４．発表年
2019年

１．発表者名
Kazuya Ueki and Takayuki Hori

２．発表標題
Zero-shot Video Retrieval using a Large-scale Video Database

３．学会等名
the International Conference on Computer Vision (ICCV2019) MDALC Workshop（国際学会）

４．発表年
2019年

１．発表者名
植木 一也，堀 隆之，小林 哲則

２．発表標題
クエリ文を用いた映像検索手法の比較検証

３．学会等名
第22回 画像の認識・理解シンポジウム (MIRU2019)

４．発表年
2019年

１．発表者名
植木 一也

２．発表標題
ゼロショット映像検索のための潜在的なコンセプトの抽出

３．学会等名
画像の認識・理解シンポジウム (MIRU2018)

４．発表年
2018年

１．発表者名
Kazuya Ueki, Koji Hirakawa, Kotaro Kikuchi, and Tetsunori Kobayash

２．発表標題
Fine-grained Video Retrieval using Query Phrases - Waseda_Meisei TRECVID 2017 AVS System -

３．学会等名
International Conference on Pattern Recognition (ICPR2018)

４．発表年
2018年

１．発表者名
Kazuya Ueki, Yu Nakagome, Koji Hirakawa, Kotaro Kikuchi, Yoshihiko Hayashi, Tetsuji Ogawa, Tetsunori Kobayashi

２．発表標題
Waseda_Meisei at TRECVID 2018: Fully-automatic Ad-hoc Video Search

３．学会等名
TRECVID 2018 Workshop

４．発表年
2018年

１．発表者名
Kazuya Ueki

２．発表標題
Latent Concept Extraction for Zero-shot Video Retrieval

３．学会等名
Image and Vision Computing New Zealand (IVCNZ2018)

４．発表年
2018年

１．発表者名
植木 一也, 中込 優, 平川 幸司, 菊池 康太郎, 林 良彦, 小川 哲司, 小林 哲則

２．発表標題
クエリ文によるゼロショット映像検索 - TRECVID 2018 AVSタスクの成果報告 -

３．学会等名
動的画像処理実用化ワークショップ(DIA2019)

４．発表年
2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関