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

平成　３０　年　　６　月　　５　日現在

機関番号：　３２６８５
研究種目：　基盤研究(C)（一般）
研究期間：　2015～2017
課題番号：　１５Ｋ００２４９

研究課題名（和文）映像認識に有効な多層の識別的構造を持つ新しいモーション特徴の研究

研究課題名（英文）Research on new multi-layer motion features effective for video recognition

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）　　3,600,000円

研究成果の概要（和文）：当初の計画していた、多層の構造を持つ識別的かつ高精度なモーション特徴の抽出については、勾配の特徴や動きを捉えるオプティカルフローの特徴を抽出し、畳み込みニューラルネットワークで学習するという仕組みを構築することで、予定よりも前倒しで映像検索システムの構築を実現することができた．また作成したシステムを，米国国立標準技術研究所が主催している国際競争型映像検索・評価ベンチマーク（TRECVID）における大規模映像データベースで評価を行い，その有効性を確認した．

研究成果の概要（英文）：With regard to the extraction of discriminative and highly accurate motion features with multi-layered structure, we were able to construct video retrieval system ahead of schedule by extracting the gradient motion features, specifically the optical flow features, and training them with convolutional neural networks. In addition, we evaluated our system with a large-scale video database at the TREC Video Retrieval Evaluation benchmark (TRECVID) organized by the National Institute of Standards and Technology and confirmed its effectiveness.

研究分野：　知覚情報処理

キーワード：　映像認識　映像検索　モーション認識　多層　CNN　TRECVID

１．研究開始当初の背景

一般物体認識などの静止画からの画像のカテゴリ識別技術は，2004 年頃から盛んに研究が行われ，ヒストグラムの勾配を特徴とする SIFT 特徴量と，ヒストグラムを量子化する Bag-of-Features（BoF）を使った手法の利用により，実用的な精度になってきた．一方，一般物体認識のコンテスト ILSVRC2012 において，Deep learning の一種である Convolutional Neural Network （以下，CNN）を使った手法を用いたチームが，SIFT＋BoF の延長線上の手法と比べて，圧倒的な精度を叩き出した．映像認識の分野においても，各フレームを静止画と考え，静止画で学習された CNN を適用することで，比較的高い認識精度が得られると報告されていた．しかしながら，静止画用の CNN では，映像から動作やイベント等のモーションを直接的に認識することができなという問題点があった．

一方，ある時間区分や画像領域での特徴の出現頻度を量子化する Dense trajectory というモーション特徴が最も良いとされていた（Heng Wang et al., International Journal of Computer Vision, 2013）．申請者らは，SIFT＋BoF や Dense trajectory といったヒストグラムの量子化による手法と，CNN を使った手法の関連性に着目し，両者の優れた機能を取り入れることで，一般物体認識や映像解析の精度が向上できると考えた．特に以下の 3 点が前者の手法における問題点と考えた．

① 特徴抽出の段階でクラスラベルが考慮されていない（識別的な要素が入っていない）
② ローカルな領域における位置関係が利用されていない
③ 多層の構造の中で小さな領域から大きな領域まで段階的に考慮する仕組みがない

２．研究の目的

映像から特定の意味的な特徴を抽出してタグ付けを行うこと，大量の映像から必要となるシーンを検索ワードから検出することを実現するため，多層の構造を持つ，識別に有効な新しいモーション特徴の抽出を研究の目的とした．特に，静止画の認識において高精度を達成している CNN の考え方を，モーション特徴にも適応し，多層の構造を持つ新しい識別的な特徴抽出方法を開発する．また同時に，モーションを認識可能とする新しい CNN の構築にも取り組む．最終的に，実用化を想定した大規模映像データベースを用いて評価を実施し，映像への高精度のタグ付け，大量映像からの検索を可能とすることを目的とした．

３．研究の方法

大量の映像から特定の意味的な特徴やシーンを高精度で検出する機能を実現するため，本研究計画では以下の２つの方向性で研究項目を実施する予定であった．

① 多層の構造を持つ識別的かつ高精度な新しいモーション特徴量の開発（平成２７年度）
② 時間変化を捉える機能を持つ新しいモーション認識用 CNN の作成（平成２８～２９年度）

最終的な目的は，大規模映像データから特定の意味的な特徴を検出することのため，評評価に使用するデータベースについては，米国国立標準技術研究所（NIST）が主催している国際競争型映像検索・評価ワークショップ（TRECVID）で使用する 100 万を越える映像データを使用して研究を遂行した．

４．研究成果

平成２７年度は，多層の構造を持つ識別的かつ高精度なモーション特徴量の研究を行った．映像から動作や動きなどのイベントを認識するための従来手法である Dense trajectory の考え方を，様々なタスクにおいて高い識別精度を示している CNN に融合することで，新しい識別的なモーション特徴を開発することを目指した．具体的には，Dense trajectory で使われる勾配の特徴や動きを捉えるオプティカルフローの特徴を抽出するため，映像の各フレームから勾配画像とオプティカルフロー画像を作成し，それらを CNN で学習するアプローチを取った．このように学習した勾配画像用 CNN に勾配画像を，オプティカルフロー画像用 CNN にオプティカルフロー画像を入力したのち，各 CNN の中間層から識別的な特徴を抽出した．特徴が抽出された後は，カテゴリ毎に勾配画像用の識別器と，オプティカルフロー画像用の識別器を作成し，最終的に複数識別器の統合を行った．

本手法の有効性を確認するため，TRECVID ベンチマークで用いられる大規模映像コーパス（2014 年の学習映像とテスト映像）を用いて評価を行った．評価指標には，各カテゴリの適合率の平均（mean Average Precision：mAP）という TRECVID で使用されている指標を用い，元画像から直接特徴を抽出する従来方法と相補的な特徴が得られることが確認できた．具体的には，ImageNet 画像データベースで学習された CNN を特徴抽出に用いた従来手法の mAP が 28.49 だったのに対し，本研究提案のモーション特徴量を加えることにより 30.97 まで向上することが確認できた．

平成２８年度から平成２９年度にかけて実施を予定していた「時間変化を捉える機能を持つ新しいモーション認識用の CNN の作成」については，平成２７年度の後半から前倒しで研究を遂行でき，大規模な映像データベースの中から，人物の特定の動作を含んでいる映像を高精度で検索可能なシステムを

構築することができた．単純な動作については，単語ベースのクエリを入力することで検索が可能となったが，一方で，一緒に写り込んでいる物体や，周りのシーンも同時に考慮した，より複雑な動作については検索することが困難であることもわかった．そのため，平成２８年度の途中からは，単純な動作の検出に加え，一緒に写り込んでいる物体や，周りのシーンも同時に考慮し，複数の条件をすべて満たすようなクエリ文を用いた複雑な動作の認識という，より難しい課題にも取り組んだ．これにより，対象となる学習データが全く存在しないような，ゼロショットでの環境下であっても，「ダイビング用のウェットスーツを着たダイバーが，水中で泳いでいる」，「群衆が街の大通りで夜間にデモ行進を行なっている」，「男の人が屋外でギターを弾いている」等，複数の条件を含んだクエリ文から映像を検索することが可能となった．クエリ文による映像検索結果を図１，２に示す．

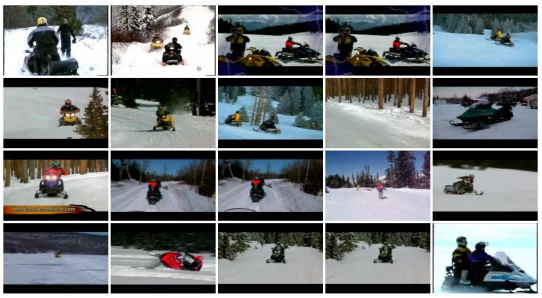

図１．クエリ文「one or more people driving snowmobiles in the snow」に対する映像の検索結果

図２．クエリ文「one or more people swimming in a swimming pool」に対する映像の検索結果

また，研究を遂行する中で，下記の２つの知見を得ることができた．１つ目は，クエリ文中のキーワードに対応するクラスのカバー率を高めるため，様々な画像・映像データセットで学習された物体・人・シーン・動作等のコンセプト識別器を大量に準備し，その組み合わせによりクエリ文を表現することの重要性である．２つ目は，キーワードに対応するコンセプト識別器を選ぶ際，該当するコンセプトが見つからない場合でも，自然言語処理の手法を取り入れることで，より多くのコンセプト識別器を選択できる仕組みの有効性である．この成果を含んだシステムをTRECVIDベンチマークのAd-hoc Video Search（AVS）タスクに提出したところ，２年連続で世界１位の映像検索精度を達成し，本手法の有効性が確認できた．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１件）

① 植木 一也,“映像検索におけるディープラーニング，”日本神経回路学会誌，vol.24，no.1，pp.1-14，2017.

〔学会発表〕（計１５件）

① 平川 幸司， 菊池 康太郎，植木 一也，林 良彦，小林 哲則,“クエリ中の単語の語義絞り込みによる動画検索精度の向上，”言語処理学会 第 24 回年次大会(NLP2018)，岡山，2018.
② 植木 一也，平川 幸司，菊池 康太郎，小林 哲則,“クエリ文を用いた詳細映像検索 -TRECVID 2017 AVS タスクの成果報告-，”動的画像処理実用化ワークショップ(DIA2018)，名古屋，愛知，2018.
③ Kazuya Ueki，Koji Hirakawa，Kotaro Kikuchi，Tetsuji Ogawa，Tetsunori Kobayashi，“Waseda_Meisei at TRECVID 2017: Ad-hoc Video Search，”Notebook paper of the TRECVID 2017 Workshop，Gaithersburg，MD，USA，2017.
④ 植木 一也,菊池 康太郎,小林 哲則,“クエリ文からの映像検索 －TRECVID 2016 AVS タスクに向けた取り組み－，”動的画像処理実用化ワークショップ(DIA2017)，松江，島根，2017.
⑤ Kazuya Ueki，Tetsunori Kobayashi，“Object Detection Oriented Feature Pooling for Video Semantic Indexing，”Proceedings of the 12th International Conference on Computer Vision Theory and Applications (VISAPP 2017)，Porto，Portugal，2017.
⑥ Kazuya Ueki，Kotaro Kikuchi，Susumu Saito，Tetsunori Kobayashi，“Video Semantic Indexing using Object Detector，”Proceedings of the 15th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Application in Industry (VRCAI 2016)，Zhuhai，China，2016.
⑦ Kazuya Ueki，Kotaro Kikuchi，Susumu Saito，Tetsunori Kobayashi，“Waseda at TRECVID 2016: Ad-hoc video search，”Notebook paper of the TRECVID 2016 Workshop，Gaithersburg，MD，USA，2016.

⑧ Kotaro Kikuchi, Kazuya Ueki, Susumu Saito, Tetsunori Kobayashi, “Waseda at TRECVID 2016: Fully-automatic Ad-hoc Video Search,” Notebook paper of the TRECVID 2016 Workshop, Gaithersburg, MD, USA, 2016.
⑨ Kotaro Kikuchi, Kazuya Ueki, Tetsuji Ogawa, Tetsunori Kobayashi, “Video Semantic Indexing using Object Detection-Derived Features,” Proceedings of the European Signal Processing Conference (EUSIPCO2016), Hungary, Budapest, 2016.
⑩ 植木 一也, 小林 哲則, “物体検出器を用いた映像の意味索引付け,”画像の認識・理解シンポジウム (MIRU2016), 浜松, 静岡, 2016.
⑪ 植木 一也, 小林 哲則, “TRECVID 2015: 映像の意味索引付けの高精度化に向けた施策,”動的画像処理実用化ワークショップ(DIA2016), 盛岡, 岩手, 2016.
⑫ 菊池 康太郎, 植木 一也, 小林 哲則, “Faster R-CNN を用いた動画像の意味索引付け,”ビジョン技術の実利用ワークショップ(ViEW2015), 横浜, 神奈川, 2015.
⑬ Kazuya Ueki, Tetsunori Kobayashi, “Waseda at TRECVID 2015: Semantic Indexing,” Notebook paper of the TRECVID 2015 Workshop, Gaithersburg, MD, USA, Nov. 16-18, 2015.
⑭ Kazuya Ueki , Tetsunori Kobayashi, “Multi-layer Feature Extractions for Image Classification - Knowledge from Deep CNNs -,” Proceedings of the 22nd International Conference on Systems, Signals and Image Processing (IWSSIP2015), pp,9-12, London, U.K, 2015.
⑮ 植木 一也, 張 雪峰, 小林 哲則, “畳み込みニューラルネットによる映像の意味インデキシング - TRECVID 2015 での試み - ,”第 18 回画像の認識・理解シンポジウム(MIRU2015), 大阪, Jul. 27-30, 2015.

〔その他〕
ホームページ等
https://kenkyu.hino.meisei-u.ac.jp/ueki-lab/index.html

６．研究組織
(1)研究代表者
植木　一也（UEKI, Kazuya）
明星大学・情報学部 情報学科・准教授
研究者番号：８０５８０６３８