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 ２９ 日現在

機関番号：１２６０８
研究種目：若手研究(B)
研究期間：2015～2018
課題番号：１５Ｋ１６０１９

研究課題名（和文）可変入力型深層学習による入力形式を問わない学習手法の確立とその映像認識への応用

研究課題名（英文）Multi-Input Deep Learning and Its Application to Video Recognition

研究代表者
井上　中順（Inoue, Nakamasa）

東京工業大学・情報理工学院・助教

研究者番号：１０７３３３９７

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）：本研究の主な成果は、画像とテキストデータを融合活用した映像認識手法を提案したことである。本手法は混合ガウス分布で単語ベクトルの分布を推定するものであり、単語ベクトルを用いて語彙拡張を行うことで、映像の意味的インデクシングの精度が向上することを示した。本成果に関する論文はACM Multimediaというマルチメディア情報処理分野の国際会議に採択されている。また、それに合わせて画像特徴量を効率的に算出するアルゴリズムを提案し、IEEE TPAMIというパターン認識分野の論文誌で発表を行った。これらの手法はTRECVID映像認識で評価されており、当初の目的であった映像認識システムの構築ができた。

研究成果の学術的意義や社会的意義
本研究の成果は、映像や画像を認識するための人工知能技術に関するものである。画像データとテキストデータの情報を組み合わせることで、認識精度が向上することを示した。これは映像のどの部分に何があるかを詳細に検索する次世代の検索システムに役立つ技術である。

研究成果の概要（英文）：In this project, we proposed a deep learning method for video recognition. The proposed method is based on vocabulary expansion using word vectors. Its performance is demonstrated on the TRECVID video dataset. We presented this work at ACM Multimedia.

研究分野：マルチメディア情報処理

キーワード：深層学習　映像認識

１．研究開始当初の背景

研究開始当初、画像・映像認識の分野では、多層ニューラルネットワークのパラメータを学習する深層学習(Deep Learning)が注目を浴び始めていた。その時点では、従来の画像特徴量と深層学習の組み合わせや、画像とテキストのデータを融合した学習手法の開発が研究課題として挙げられる状況であった。また、深層学習は令和元年現在ほど社会に浸透しておらず、当初は比較的新しいテーマであった。

２．研究の目的

本研究の目的は、ネットワークの構造が入力に合わせて事後的に変化する学習手法の提案し、入力の形式を問わない情報検索の基盤を構築することである。特に、画像データとテキストデータなどの異なるデータを組み合わせて利用した深層学習の有効性を示すことが目標である。評価実験では、大規模な映像データセットを用いることで、定量的な評価を実施する。

３．研究の方法

画像・テキスト・音声を研究の対象とし、画像からの特徴抽出、テキストデータの利用、音声からの特徴抽出、それらを融合した映像認識システムの構築に取り組む。評価実験は、映像認識に関する TRECVID データセットおよび画像認識に関する ImageNet データセットで行う。

４．研究成果

本研究の主な成果は下記の３点である。(1)では画像データとテキストデータを融合した映像認識手法を提案した。本成果に関する論文は ACM Multimedia というマルチメディア情報処理分野のトップカンファレンスに採択されている。(2)では画像特徴量抽出の高速化手法を提案した。本成果は IEEE TPAMI というパターン認識分野の論文誌に掲載されている。(3)では TRECVID と呼ばれる映像認識データセットへの適用を行った。本成果は TRECVID Workshop で発表している。

(1) 画像データとテキストデータを融合した映像認識

本研究項目では、映像から物体・動作・シーンを検出することを目的としたセマンティックインデクシングに対し、単語ベクトルを用いた語彙拡張方法を提案した。本手法は画像データとテキストデータを融合的に利用した学習方法であり、画像から畳み込みニューラルネットワークにより抽出した特徴量と事前に学習された単語ベクトルを組み合わせて映像認識を行う。組み合わせ方法としては、ベクトル量子化(VQ)、k-近傍探索(k-NN)、混合ガウス分布(GMM)の３つのモデルに基づいた方法を提案した。

評価実験は、TRECVID Semantic Indexingデータセットを用いて行った。データセットは100時間のインターネット映像からなり、カメラの切り替わり部分を示すショット境界の情報が与えられている。ショット総数は264,673で、学習用119,685ショットとテスト用144,988ショットに分けられている。タスクは346種類の物体・動作・シーンのからなる意味的概念の検出であり、Boat, Airplane, Dancing, Walkingなどのカテゴリが含まれている。評価尺度はMean Average Precisionである。

評価結果は表１の通りで３種類のモデルに基づいた提案手法が、当時画像認識手法として広く使われていたFisher Vectorに基づく手法を上回っていることが分かる。また、意味的概念ごとの認識率は図１の通りで提案手法が車などの物体の学習に特に有効であることが分かった。一方、人物の動作は認識精度が低く、今後、時系列のモデル化が必要になることが明らかとなった。

表１．認識精度の比較

手法	Mean AP
VQ	12.7
K-NN (k=10)	14.0
GMM	15.3
FV (従来手法)	10.1

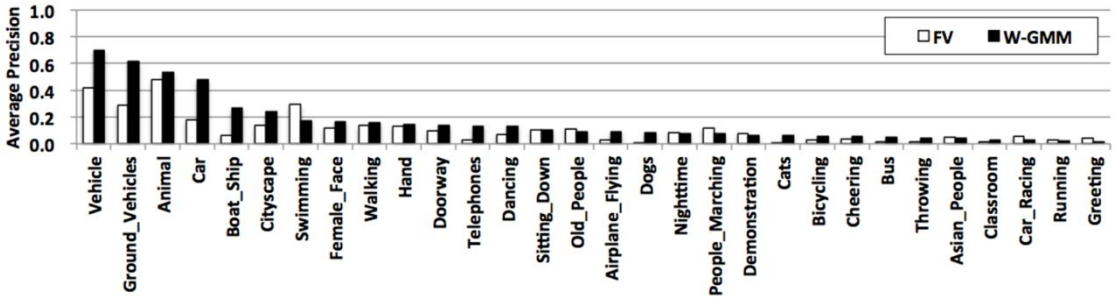


図 1　意味的概念ごとの認識率

(2) 画像特徴量抽出の高速化手法

本研究項目では、画像のサイズに依らない画像特徴量の抽出に着目し、その量子化や分布推定の高速化手法を提案した。特徴量は畳み込みニューラルネットワークやその他の局所特徴量に基づいたものである。提案のアルゴリズムは、映像データにおいて、被写体の動きが少ない連続した画像フーレムで、特徴量の分布の変化が少ないことを利用し、計算量を削減するものである。

主な実験結果は図２の通りで、画像のある点とその周囲から特徴量を抽出し、それらの量子化を行う際の計算量が平均93.0%削減可能であることを示した。本手法は画像データにも有効であるが、映像データに対しても計算量削減の効果が高いことを示すことができた。

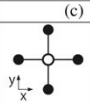
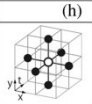
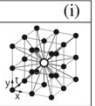
	(c)	(h)	(i)
Neighbors $B(x)$			
$\|B(x)\|$	4	6	26
$\Delta_{i,i+1}$	116.2	82.0	101.5
Time (msec)	1542.9	1267.3	1512.9
Reduction (%)	91.5	**93.0**	91.6

図 2　近傍特徴量と高速化結果

(3) TRECVID映像データセットへの適用

最後に、TRECVID と呼ばれる米国標準技術研究所が主催の国際競争型ワークショップで提供されている、マルチメディアイベント検出およびアクティビティ検出データセットへの適用実験を実施した。前者は、複数の物体や人物が関連した事象（イベント）を映像データから検出するタスクであり、(1)の手法を適用した結果、楽器など幾つかの物体に関して検出精度が向上したが、平均的には大きな精度向上は見られなかった。後者は、監視カメラの映像から人と車に関連したコミュニケーションの発生を検出するもので、人物位置や物体位置の特定に関する前処理が不十分であり、現在も研究を継続中である。

以上の研究で、当初の目的の１つであった、画像とテキストデータを融合的に活用した学習手法と、それに基づいた映像認識システムの構築ができた。しかし、音声に関しては、予想通りの効果が得られていない。これは音を学習するためのラベル付きデータが不足していたことが原因であるが、本研究内では、音データにラベルを付けてデータセットを作成する費用が確保できなかったため、今後何らかの形で研究を継続展開したい。

５．主な発表論文等

〔雑誌論文〕（計 2 件，査読あり）

Nakamasa Inoue, and Koichi Shinoda, “Fast Coding of Feature Vectors using Neighbor-To-Neighbor Search,” IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 38, no. 6, pp. 1170-1184, 2015.

Nakamasa Inoue, and Koichi Shinoda, “Semantic Indexing for Large-Scale Video Retrieval,” ITE Transactions on Media Technology and Applications, vol. 4, no. 3, pp. 209-217, 2016.

〔学会発表〕（計 3 件）

Nakamasa Inoue and Koichi Shinoda, “Vocabulary Expansion Using Word Vectors for Video Semantic Indexing,” Proc. ACM Multimedia, pp. 851-854, 2015.

井上 中順, 篠田 浩一, “単語ベクトルによる語彙拡張を用いた映像のセマンティックインデクシング,” 電子情報通信学会技術研究報告 PRMU, vol. 115, no. 388, pp. 75-80, 2015

Nakamasa Inoue, Chihiro Shiraishi, Aleksandr Drozd, Koichi Shinoda, Shi-wook Lee, and Alex Chichung Kot, Activity Detection in Extended Video using Action Tubelets (VANT at TRECVID 2018), Proc. TRECVID workshop, 2018.

〔図書〕（計 0 件）

〔産業財産権〕

○出願状況（計 0 件）

○取得状況（計 0 件）

〔その他〕

ホームページ等

６．研究組織

(1)研究分担者：なし

(2)研究協力者：なし