第四人称キャプショニング：相補性を有する分散視覚を用いたヒト－ロボット共生空間の状況記述

中嶋 一斗[1,a)]　岩下 友美[2,b)]　河村 晃宏[3,c)]　倉爪 亮[3,d)]

1. はじめに

高齢者人口の増加が社会問題となりつつあり，病院や介護現場における要介護者数の増大や労働者人口減少に伴う医療従事者の人手不足が深刻である．そのため，ヒトに代わって要介護者の生活支援を行う共生型サービスロボットの需要が高まっている．特に，ロボット単体では限界のあるセンシング能力を補うべく，共生空間全体にセンサを分散配置することで空間そのものに知性を与える「空間知能化」が注目されている [1]．空間知能化の主目的はロボットの作業効率・知能を補助することであるが，居住者や保護者などのヒトの認知能力・記憶能力を高める仕組みとしても有用である．例えば，居住者の日常行動を複数分散センサを用いて自動かつ定量的に記録できれば，生活習慣の改善や遠隔地からの見守り・安否確認などに利用できる．そこで，本研究では，ヒトやロボットの振る舞いから周辺環境の様子に至るまで，知能化空間における日常生活の自動要約を目的とし，分散視覚から得られる画像群から自然言語による状況記述を行うためのアーキテクチャを提案する．

2. 関連研究

2.1 ヒト－ロボット共生空間の状況記述

状況記述に関するこれまでの研究は，多くがヒトの振る舞いのみに注目し，計算機にとっての可読性・処理効率が高いオントロジを用いて形式的に表現するのが一般的であった．例えば，森ら [5] は，日常生活の要約を目的として，知能化空間で計測されるヒトの位置情報から 5 種類の行動クラスに分類している．これらは，特徴的な生活パターンの解析や異常行動の検知などに応用できるが，提示された状況をヒトが自然に解釈できない問題がある．

1 九州大学大学院システム情報科学府
2 カリフォルニア工科大学ジェット推進研究所
3 九州大学大学院システム情報科学研究院
a) k_nakashima@irvs.ait.kyushu-u.ac.jp
b) Yumi.Iwashita@jpl.nasa.gov
c) kawamura@ait.kyushu-u.ac.jp
d) kurazume@ait.kyushu-u.ac.jp

図 1: 第四人称視点の概念図（小説の例）

2.2 画像キャプショニング

近年の深層学習手法の発展により，異なるモダリティ間の複雑な写像を統一的なモデルで表現することが可能になった．その成果の一つとして，与えられた画像の内容を自然言語で説明する画像キャプショニングが高精度に実現している [4]．しかし，この技術を知能化空間で実応用しようとした場合，配置したカメラの視野や解像度の問題により，単一の視点では正確な状況記述に充分な視覚特徴を抽出できない可能性が高い．

2.3 相補性のある分散視覚「第四人称視点」

著者らは知能化空間における分散視覚の人称性に着目し，相補性を有する 3 つの人称視点に絞って融合することで正確な情報抽出を図る「第四人称視点」を提案している [2]．具体的には，知能化空間の分散カメラから得られる大量の視覚情報のうち，ウェアラブルカメラなどから得られる居住者の視点を一人称視点とすると，ロボット搭載のカメラを二人称視点，環境固定型カメラを三人称視点と定義できる．これら 3 つの人称視点を相補的に組み合わせることができれば，異なる 3 つのモダリティを包含した全く新しい視点「第四人称視点」を構築することができる（図 1：読者の視点）．本稿では，代表的な画像キャプショニングアーキテクチャを 3 入力に拡張し，知能化空間から得られる相補性を有する 3 つの人称視点画像を用いたヒト－ロボット共生空間の状況記述手法，すなわち第四人称キャプショニングを提案する．

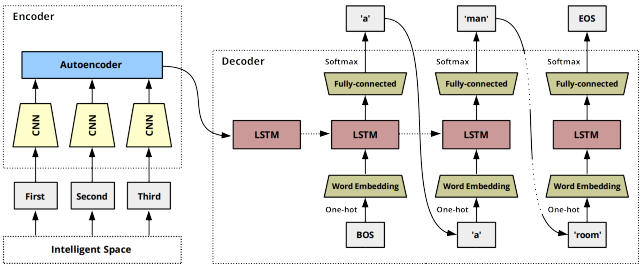

図 2: 提案モデルのアーキテクチャ

3. 第四人称キャプショニング

3.1 概要

これまでの代表的な画像キャプショニングアーキテクチャは，単視点から得た対象画像をコンパクトな特徴ベクトルに符号化する畳込みニューラルネットワーク (CNN) と，その特徴ベクトルから連鎖する単語列を順次予測するための回帰型ニューラルネットワーク（RNN）をシームレスに接続することで，明示的に構文規則を定義する必要なく，教師データのみから End-to-end で学習可能なアーキテクチャを構築している．そこで，本稿では，知能化空間における 3 つの人称視点画像から抽出される CNN 特徴を融合し，RNN に入力することで単語系列を生成するアーキテクチャを提案する（図 2）．

3.2 入力画像の符号化と融合

各人称視点の画像は，まず学習済み CNN を用いて特徴ベクトルに変換する．本研究では，ImageNet で学習した VGG-19 モデル [3] を利用して，4096 次元の CNN 特徴を得る．また，提案アーキテクチャを End-to-end で学習しようとした場合，視覚特徴の融合過程を獲得するために，3 つの人称視点画像が一組となった大規模な教師データが必要である．そこで，本研究では第四人称視点を設置した模擬生活空間上で教師なし多視点画像を収集し，それらから抽出される CNN 特徴に基いて積層自己符号化器（Stacked Autoencoder）を事前構築することで，データセットの新規構築を必要としない視覚特徴の融合を行う．

3.3 単語系列への復号化

積層自己符号化器により融合された第四人称視点の視覚特徴は，まず RNN に入力することで，隠れ変数を初期化する．次に，文の開始を表す単語 BOS を語彙数 K に基づく 1-of-K 表現に変換し，RNN に入力することで，RNN からは連鎖する単語の確率分布が出力される．以降のステップからは，出力単語を再度 RNN に入力し，文の終端を表す単語 EOS が出現するまで同様に繰り返す．本研究では，文献 [4] と同様に 1 層の LSTM-RNN を利用する．

(a) a man standing in a kitchen preparing food

(b) a white refrigerator freezer sitting inside of a kitchen

図 3: 提案手法による生成例（左から一人称，二人称，三人称視点）

4. 実験

4.1 実験方法

まず，第四人称視点を設置した模擬生活空間上で，居住者の物品探索行動やロボットへの受け渡し等を撮影し，第四人称視点データセットを構築した．この第四人称データセットを利用して，視覚特徴を融合する積層自己符号化器の教師なし学習を行う．

次に，入力画像情報に基いて単語の連鎖過程をモデル化する LSTM-RNN の教師あり学習を行う．ここでは，CNN および積層自己符号化器のパラメータを固定し，LSTM-RNN の内部パラメータのみを学習対象とする．特に，単視点を対象とする大規模公開データセットを有効利用するために，各人称視点毎に画像と説明文の組を与え，学習する．この時，対象外の人称視点に対応する初段の層結合をカットし，学習後に 3 つの人称視点画像を用いて推論する場合は初段の全出力値を 1/3 倍する．

4.2 生成結果

図 3 に提案アーキテクチャによる説明文の生成例を示す．画像内容に即した説明文生成には至っていないが，3 つの視点いずれかの視覚特徴を反映している傾向が見られた．

5. まとめと今後の展望

本稿では，ヒト－ロボット共生空間における日常生活の自然言語による自動描写を目的とし，第四人称視点から得られる 3 つの人称視点画像に基づく自然言語文生成手法，第四人称キャプショニングを提案した．さらに，ヒト－ロボット共生空間の実証実験施設で実際に撮影した 3 つの人称視点画像を用いて，提案手法による説明文生成実験を行った．実験の結果，入力した 3 つの視点いずれかの画像内容が生成文に反映される傾向が見られた．今後は，3 視点の画像組に対応した空間説明文付きデータセットの新規構築を行い，End-to-end 学習および提案手法の定量評価を行う．

謝辞

本研究は文部科学省科学研究費補助金基盤研究（A）（課題番号 26249029）の支援を受けた.

参考文献

[1] Kurazume, R., Pyo, Y., Nakashima, K., Tsuji, T. and Kawamura, A.: Feasibility study of IoRT platform "Big Sensor Box", *The IEEE International Conference on Robotics and Automation*, pp. 3664–3671 (2017).

[2] Nakashima, K., Iwashita, Y., Pyo, Y., Takamine, A. and Kurazume, R.: Fourth-Person Sensing for a Service Robot, *The IEEE International Conference on Sensors*, pp. pp.1110–1113 (2015).

[3] Simonyan, K. and Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition, *CoRR*, Vol. abs/1409.1556 (2014).

[4] Vinyals, O., Toshev, A., Bengio, S. and Erhan, D.: Show and Tell: A Neural Image Caption Generator, *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3156–3164 (2015).

[5] 森　武俊, 野口博史, 佐藤知正：センシングルーム：部屋型日常行動計測蓄積環境第 2 世代ロボティックルーム，日本ロボット学会誌 = Journal of Robotics Society of Japan, Vol. 23, No. 6, pp. 665–669 (2005).