第四人称視点に基づく情報構造化空間の状況説明文生成

○中嶋一斗 (九州大学) 岩下友美 (NASA/Caltech JPL) 倉爪亮 (九州大学)

1. はじめに

情報構造化空間とは，人と生活支援ロボットの共生を実現するために有効な知的生活環境の枠組み [1] である．一般生活環境に多数のセンサを分散配置することで，ロボット単体の処理性能・身体性に制限されることなく，環境内を歩き回る居住者の行動や日用品位置等の広範囲で多様な環境情報を計測できる．これらの環境情報は，ロボットの計測能力を補強するだけでなく，居住者の記憶補助や内省に利用できる．例えば，情報構造化空間が多角的に捉える日常生活状況を自然言語文として定常的に描写・記録することで，文字や音声を媒介した直感的な情報検索や居住者との質問応答に応用できる．

本研究では，情報構造化空間の生活状況を説明する自然言語文の自動生成手法を提案する．特に，情報構造化空間における相補的多視点に基づくメタ視点の概念「第四人称視点」 [2, 3] に基づき，単視点推論に比べて正確な生活状況記述を試みる．本稿では，多視点画像を入力とする説明文生成のためのニューラルネットモデル，既存の大規模データセットを有効利用するための半教師あり学習手法について提案し，実際の情報構造化空間で収集した画像組から生成実験を行う．

2. 関連研究

画像・映像に基づくライフログ分野では，ユーザに装着したウェアラブルカメラから得られる一人称視点映像が重要視されてきた．例えば，Miyanishi ら [4] は，映像記録時のジェスチャから想起して過去の一人称視点映像を検索するシステムを提案している．Nakayama ら [5] は，物体キーワードを指定して，一人称視点画像を検索するウェアラブルシステムを開発している．しかし，従来のジェスチャやキーワード等を用いた要求生成は，ユーザに一定のスキル習得が求められる．将来の知的情報支援システムでは，人にとって直感的な自然言語を利用することが望ましい．

近年では，深層学習技術の進展に伴い，画像説明文の自動生成に関する取り組みが活発になってきた．その多くは，畳込みニューラルネット（Convolutional Neural Network，CNN）による画像特徴抽出と再帰型ニューラルネット（Recurrent Neural Network，RNN）による言語モデル（RNN-LM）の組み合わせにより実現される．Vinyals ら [6] は，ニューラル機械翻訳における符号化・復号化プロセスを類推し，学習済み物体認識 CNN による画像特徴の抽象化と RNN-LM の条件付け（符号化）および RNN-LM による後続単語の逐次予測（復号化）を行う end-to-end ニューラルネットを構成した．さらに，各ステップの単語予測において本質的に重要な局所画像領域/物体領域への逆伝搬経路を短縮する注意機構（attention mechanism） [7, 8] が導入されたことで，生成精度は大幅に改善した．

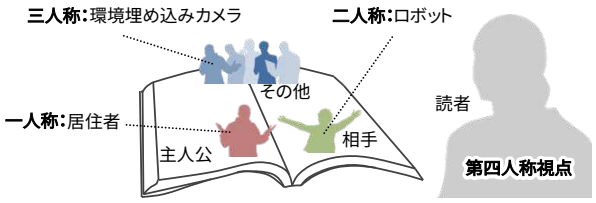


図 1: 情報構造化空間における第四人称視点の概念図

ライフログ分野での取り組みとしては，Fan ら [9] が胸部装着型カメラから得られる一人称視点画像から，装着者の状況を自然言語で記述する手法を提案している．この手法では，著者らが独自に構築した約 700 組の一人称視点画像と説明文の対を用いて符号化・復号化モデル [6] を学習している．

3. 第四人称視点

本節では，情報構造化空間の分散視覚から得られる巨視的概念について述べる．情報構造化空間において，居住者の視点を一人称視点とすると，ロボットの視点を二人称視点，環境埋め込みカメラの視点を三人称視点と定義することができる．一人称視点は，カメラ装着者の内発的動作を直接反映するため，日常生活動作や日用品の操作履歴を記録しやすいが，装着者周辺の局所的な情報しか得られない．一方，情報構造化空間におけるロボット搭載カメラ（二人称視点）や環境埋め込みカメラ（三人称視点）は，人の行動や環境との相互作用を外部視点から大局的に観測することができる．しかし，撮影対象物体までの距離が長く，解像度や死角の影響を受けやすい．そこで，固有の観測範囲や解像度を有する一・二・三人称視点カメラの視覚情報を相補的に融合することで，情報構造化空間の正確な状況理解ができると考えた．このメタ視点による計測の概念を，小説の文面から登場人物（一・二・三人称視点）の心情・関係性を読み取る読者視点のアナロジーとして，第四人称視点と定義した（図 1） [2].

4. 生成モデル

本節では，第四人称視点から定義される三種の画像から，単独の自然言語説明文を生成するニューラルネットモデルについて説明する．

4.1 画像特徴抽出と融合

まず，同期撮影された一・二・三人称視点画像を事前学習済みの物体認識 CNN に入力し，それぞれの画像特徴ベクトルを得る．次に，三視点の特徴ベクトルを結合し，共通特徴空間に線形写像する．写像後のベクトルは双曲線関数による活性化，バッチ正規化を行った．これを融合層とする．

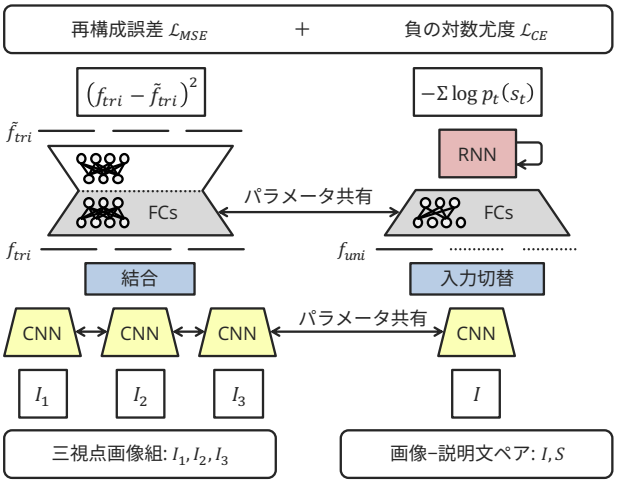


図 2: 提案手法の半教師あり学習概略図．図中の I_1, I_2, I_3 は，それぞれ一・二・三人称視点画像．$I, S = \{s_1, \ldots, s_n\}$（$s_t$ は t 番目の単語）は公開データセットに含まれる単一画像と説明文のペア．

4.2 説明文生成

Long Short-Term Memory（LSTM）により，説明文の構成単語を逐次的に予測する．使用する全ての単語 s_t は語彙ベクトルの one-hot 表現として与えられる．LSTM は各時刻 t で，入力ベクトル x_t および前の隠れ状態 h_{t-1} から，現在の隠れ状態 h_t を更新する．前節で得られた画像特徴は，最初の時刻で x_0 として与えられる．次時刻以降は，前時刻の単語 s_{t-1} を同じ次元の空間に写像後，x_t として与える．同時に，h_t から規定語彙の尤度 p_t を出力し，最大確率の単語を選択する．

5. 半教師あり学習

画像説明文生成モデルは，一般的に教示文 $S = \{s_1, \ldots, s_n\}$ の最尤推定により学習される．すなわち，次式に示す負の対数尤度を損失として最小化する．

$$\mathcal{L}_{CE} = -\sum_{t=1}^{|S|} \log p_t(s_t | I, s_1, \ldots s_{t-1}) \tag{1}$$

これは，教示文を構成する n 単語分の結合確率を最大化することと等しい．また，この損失を計算するために，前章で述べたように三視点の画像特徴を融合し，LSTM の初期入力ベクトルとして与えなければならない．しかし，従来の学習手法の下では，三視点画像組に対応する新規の説明文を教示する必要があり，MSCOCO [10] を始めとする既存の大規模データセットを有効利用することができない．そこで，三視点特徴の融合学習と言語モデルの学習を，別々の損失関数として定式化した半教師あり学習手法を提案する．提案手法の概略図を図 2 に示す．

提案手法では，情報構造化空間から収集した三視点画像組（図 2 中 I_1, I_2, I_3）による教師なし自己教示と，単一画像を対象とした既存データセット（図 2 中 I, S）による単視点文生成教示から構成される．まず，三視点画像組が与えられると，CNN により画像特徴を抽出し，結合ベクトル f_{tri} を生成する．結合ベクトルは，融合層をエンコーダとした自己符号化器を通過し，$\tilde{f}_{tri}$ として再構成するように学習する．すなわち，次式の平均二乗損失を最小化する．

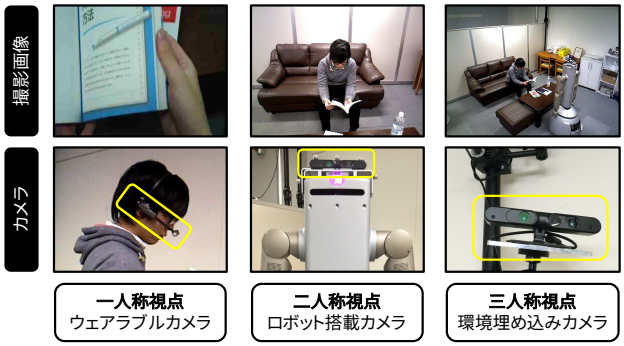


図 3: 三視点同期映像の収集環境

$$\mathcal{L}_{MSE} = (f_{tri} - \tilde{f}_{tri})^2 \tag{2}$$

教師あり単一画像を与える場合は，いずれかの視点に CNN による画像特徴を設定し，残り二つの視点を零ベクトルとした結合ベクトル f_{uni} を生成する．この時，融合層通過後のベクトルは，通常通り式 1 を最小化する．ただし，ベクトル f_{uni} の各要素は，この学習時のみ 3.0 で乗算する．最終的な損失関数 $\mathcal{L}$ は，係数 λ を用いて次の重み付き和で与えられる．

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{MSE} \tag{3}$$

6. 実験

各手法は，深層学習フレームワーク PyTorch を用いて実装し，Geforce GTX Titan X 上で実行した．また，画像特徴抽出を行う CNN として，ImageNet で事前学習した ResNet-152 [11] を利用し，Global Average Pooling から出力される 2048 次元ベクトルを抽出した．LSTM の入力層ユニット数，隠れ層ユニット数はそれぞれ 512 とした．本実験では，重み付き和損失の係数 λ を 1 とした．

6.1 データ

まず，研究室が保有する情報構造化空間実験環境において三視点同期映像の収集を行った（図 3）．実験環境は，机やソファ，テレビなどの家具を配置しており，ウェアラブルカメラを装着した居住者が一般日常行動を行いながら，一人称視点映像を記録した．また，一体のサービスロボットを空間内に配置し，常に被験者を監視する形で二人称視点映像を記録した．最後に，居住者とサービスロボットが同時に撮影できる位置にカメラを固定配置し，三視点視点映像を記録した．カメラを配置を変えながら，200 組ほどの三視点画像データを得た．

生成文教示のためのデータには，約 11 万枚の画像から構成される MSCOCO [10] を使用した．各画像には，五つの文候補が付与されている．教師データとして利用するための前処理として，句読点と記号を除去した後，全ての単語を小文字に変換した．また，文生成の始まりと終わりを表す単語<start>と<end>を用意し，出

現頻度が 5 回以下の単語は<unknown>と定義した．最終的な語彙数は，10,107 語となった．

6.2 実験結果

図 4 に，Vinyals ら [6] のモデルを用いていずれかの単視点の画像特徴を RNN に入力した場合，三視点の特徴ベクトルの平均を同モデル [6] の RNN に入力した場合，および提案モデルを用いた場合の生成例を示す．また，各生成結果においては，Grad-CAM [12] を用いて各単語に寄与する画像特徴領域を可視化した．生成された説明文はいずれも不適切な単語を含んでいるが，各視点の解像度に応じて画像内の特徴をいくらか捉えている．例えば，一人称視点入力では，人が手中で物体を把持している様子が生成されている（図 4a, “a person holding a pair of scissors in their hand”）．二人称視点入力では，人がソファに座っている様子と膝上の物体を記述している（図 4b, “a woman sitting on a couch with a cat in her lap”）．三人称視点入力では，座っている場所がどんな部屋かといった，より大局的な情報が含まれている（図 4c, “a man sitting in a chair in a living room”）．一方，これらの三視点の特徴を単純に平均した結果では，人の着座に関して記述できているものの，後続の文は実際の状況を表現できていない（図 4d, “a woman sitting on the floor playing a video game”）．それら後続単語に対応する寄与画像領域に関しては，顕著な活性領域が見られないことから，RNN-LM の統計的傾向に従ってデータセット頻出の文節を自動生成したものと考えられる．最後に，提案手法による生成結果では，居住者がソファに座っている様子と物体を使用している様子を記述できている（図 4e, “a woman sitting on a couch using a laptop computer”）．使用対象物体に関して誤認識しているが，当該単語に対応して一人称画像の本の領域に反応していることが分かる．また，“on a couch” という文節に対応して，二人称視点の三人称視点のソファ領域に同時活性化していることから共起情報を捉えているものを考えられる．一方で，場所に関する大局的情報は含まれなかった．

7. まとめと今後の予定

情報構造化空間において得られる相補的な三種の視点による，状況説明文の自動生成手法を提案した．特に，本稿では，最も基本的な画像説明文生成モデルを多視点入力に拡張し，生成方法について述べた．また，単一画像と説明文の対から構成される既存大規模データセットを有効利用するために，単視点ベースの文生成教示と三視点ベースの教師なし融合学習を組み合わせた半教師あり学習手法を提案した．情報構造化空間の三視点画像データと MSCOCO を用いた実験では，半教師あり学習で部分的に情報構造化空間の状況を記述できることを確認した．また，いずれかの視点のみを入力した場合と三視点を単純に平均した場合と比較し，いくらか視点間の情報を統合できているものの，居住者周辺の情報を中心とした二人称視点の記述内容に近い傾向が観察された．課題として，単視点入力そのものの生成精度，教師データの統計的な文長の影響が挙げられた．今後は，より高解像度な画像を使った生成実験，空間注意機構 [7] の導入，および定量的評価を行うためのデータセット構築を行う予定である．

謝辞

本研究は，JST, CREST, JPMJCR17A5 の支援を受けたものである．また一部は文部科学省科学研究費補助金基盤研究（A）（課題番号 JP26249029）の支援を受けた．

参 考 文 献

[1] 倉爪亮, ピョユンソク, 辻徳生, 河村晃宏: “情報構造化環境プラットフォーム ROS-TMS と Big Sensor Box の提案”, 日本ロボット学会誌, vol.35, no.4, pp.346–357, 2017.

[2] K. Nakashima, Y. Iwashita, Y. Pyo, A. Takamine and R. Kurazume: “Fourth-Person Sensing for a Service Robot”, *Proc. of the IEEE Conf. on Sensors*, pp.1110–1113, 2015.

[3] 中嶋一斗, 岩下友美, ピョユンソク, 高嶺朝理, 倉爪亮: “サービスロボットのための第 4 人称センシングの提案”, 画像の認識理解シンポジウム *(MIRU2015)*, SS4-11, 2015.

[4] T. Miyanishi, J. Hirayama, Q. Kong, T. Maekawa, H. Moriya and T. Suyama: “Egocentric Video Search via Physical Interactionse”, *Proc. of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI)*, pp.330–336, 2016.

[5] H. Nakayama, T. Harada and Y. Kuniyoshi: “Ai goggles: Real-time description and retrieval in the real world with online learning”, *Canadian Conference on Computer and Robot Vision (CRV)*, pp.184–191, 2009.

[6] O. Vinyals, A. Toshev, S. Bengio and D. Erhan: “Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge”, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol.39, no.4, pp.652–663, 2017.

[7] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudino, R. Zemel and Y. Bengio: “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”, *Proc. Int. Conf. on Machine Learning (ICML)*, pp.2048–2057, 2015.

[8] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould and L. Zhang: “Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering”, *arXiv preprint arXiv:1707.07998*, 2017.

[9] C. Fan and D. Crandall: “DeepDiary: Automatically Captioning Lifelogging Image Streams”, *Proc. European Conf. on Computer Vision International Workshop on Egocentric Perception, Interaction, and Computing (EPIC)*, pp.459–473, 2016.

[10] X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollár and C. L. Zitnick: “Microsoft COCO Captions: Data Collection and Evaluation Server”, *arXiv preprint arXiv:1504.00325*, 2015.

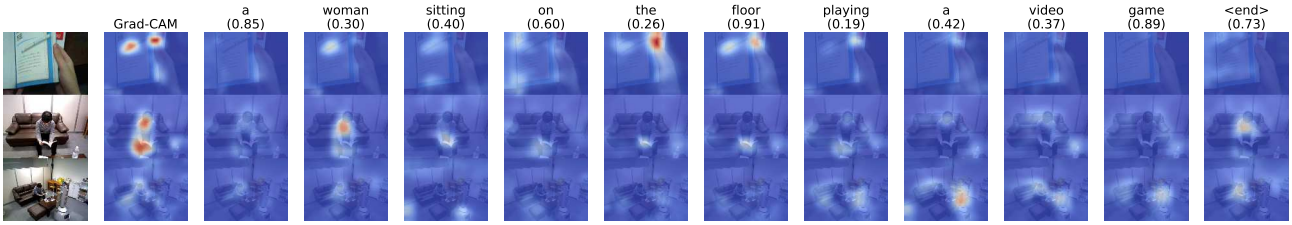
[11] K. He and X. Zhang and S. Ren and J. Sun: “Deep Residual Learning for Image Recognition”, *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp.770–778, 2015.

[12] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh and D. Batra: “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization”, *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pp.618–626, 2017.

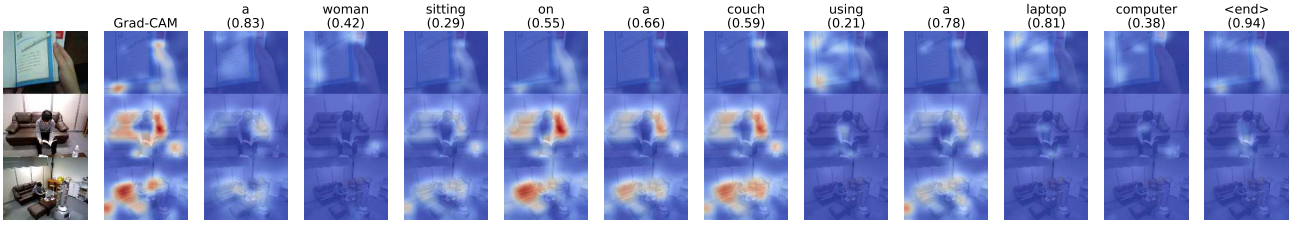
(a) 一人称単視点入力

(b) 二人称単視点入力

(c) 三人称単視点入力



(d) 一・二・三人称視点の平均画像特徴



(e) 一・二・三人称視点の融合特徴（**提案手法**）

図 4: 単視点入力，三視点平均，提案手法による状況説明文の生成結果と寄与画像領域