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
第 2 報　顕著領域クラスタリングによる視点間アテンションの統合

○中嶋一斗　倉爪亮（九州大学）

1. はじめに

近年，人の頭部または胸部に装着する小型ウェアラブルカメラが広く普及し，日常の視覚体験を一人称視点映像として容易に記録できるようになった．一人称視点映像にはカメラ装着者の注目領域や接触物体が撮影されるため，ライフログとして長期的に記録・ラベリングすることで日常行動や社会活動のパターンを分析することができる [1]．また，映像あるいは画像から抽出したラベル情報から索引付けを行うことで，ライフログユーザは膨大な記録映像群から任意シーンを検索し，閲覧することができる [1]．特に最近では，属性ごとに独立したラベリングではなく自然言語説明文（キャプション）として統一的に記述する試みが進んでいる [2]．一方で，一人称視点映像はカメラ自己運動による被写体ブレや不明瞭な映像区間を多く含むことが欠点として挙げられる．そこで本研究では，人とロボットが共存する一般生活環境を適用対象として，複数視点を用いたライフログ生成手法を開発し，正確な状況記述のための視点構成について検討する．一般生活環境における人・ロボット共生に有効なアプローチの一つとして，情報構造化空間（知能化空間）が挙げられる [3]．情報構造化空間は，一般生活環境に分散センサネットワークを構築することで，ロボット単体の処理性能や計測範囲に制限されることなく，環境全体の人の行動や物品位置等を計測する枠組みである．この枠組みでは，人の一人称視点だけではなく，ロボットカメラ（二人称視点）や環境埋込みカメラ（三人称視点）による客観視点を利用することができる．本稿では，一人称・二人称・三人称視点によるライフログ画像から情報構造化空間の状況説明文を自動生成する手法を提案する．特に，一人称視点の主体となる人の行動に焦点を当てて，検証実験を行う．以降では，複数画像を入力とする説明文生成モデル，提案手法を検証するために構築したデータセット，データセットを用いた定量的評価実験について述べ，提案手法の有効性を示す．

2. 提案手法

2.1 情報構造化空間における第四人称視点

情報構造化空間において，居住者のウェアラブルカメラから得られる視点を一人称視点とすると，ロボットに搭載したカメラを二人称視点，人・ロボットを含めた空間全体を捉える環境埋込みカメラ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，ロボット搭載カメラや環境埋め込みカメラは，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．本研究では，各視点固有の観測範囲・解像度を手がかりにより正確な状況記述が実現できると考え，この三視点を組み合わせた巨視的概念を，小説の文面から登場人物（一人称・二人称・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ている．

2.2 ベースライン

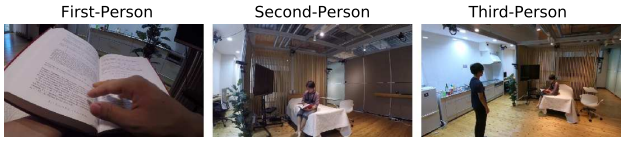
単一画像のキャプション生成分野では，学習済み物体認識 CNN による画像特徴の抽象化と，画像特徴による RNN 言語モデルの条件付けおよび後続単語の逐次予測を行うエンコーダ・デコーダ構成が主流である．特に，画像特徴を空間的に細分化し，各ステップの単語予測において適応的に選択するアテンション機構が導入されたことで，キャプションの生成品質は大幅に向上した [4, 5]．本研究では，前景・背景物体による顕著領域をアテンション候補として用いた Anderson ら [5] の `UpDown` モデルをベースラインとし，次章で複数視点入力への適応について述べる．`UpDown` モデルは，顕著領域検出器とキャプション生成器から構成される．顕著領域検出器は，初めに画像内の顕著領域を多数検出し，各領域について物体クラスおよび属性クラスを同時識別する．次に，非最大値抑制処理および識別された物体クラスの推定尤度に従って最大 N 個の領域を選別し，各領域に対応する中間表現ベクトルの組をアテンション候補とする．キャプション生成器では，単語を一つ生成するごとに前時刻の RNN 状態ベクトルに基づいてアテンション候補の重みを決定し，凸結合されたアテンション候補を RNN に入力することで規定語彙の確率分布を推定する．

2.3 顕著領域クラスタリングの導入

`UpDown` モデルでは，任意数のアテンション候補を入力することができるため，RNN の前時刻状態に基づくアテンション機構のみで複数視点間の類似特徴を同時選択，あるいは視点固有の特徴を適宜選択できる可能性がある．入力画像のみを変更する本手法を `Ensemble` と表記する．一方で，領域特徴のどの次元に着目するかは，学習されたアテンション機構に依存するため，視点ごとの見え方・コンテキストによって別のインスタンスとして表象する可能性がある．そこで本研究では，複数視点の顕著領域ベクトルを予めクラスタリングし，各クラスタの代表ベクトルを新たなアテンション候補として再編成する顕著領域クラスタリングを提案する．本稿では，k-means アルゴリズムを採用し，本手法を `KMeans` と表記する．また，本稿ではクラスタ数を 32 とした結果のみ報告する．

表 1　B-n，R，M，C，S はそれぞれ，BLEU-n，ROUGE-L，METEOR，CIDEr-D，SPICE を表す．

一人称	二人称	三人称	手法	B-1	B-2	B-3	B-4	R	M	C	S
✓			UpDown [5]	51.20	33.47	20.41	11.25	38.85	17.45	21.44	12.19
	✓		UpDown [5]	60.86	43.24	31.12	21.19	45.60	19.46	16.94	12.08
		✓	UpDown [5]	42.80	26.56	16.17	9.70	31.34	13.73	6.79	6.28
	✓	✓	Ensemble	59.14	41.97	30.45	21.06	44.09	19.13	15.18	11.40
	✓	✓	KMeans	62.31	45.34	33.16	22.91	46.22	20.19	17.76	12.21
✓		✓	Ensemble	59.06	42.78	30.47	20.28	45.16	20.33	27.71	14.37
✓		✓	KMeans	60.83	44.71	32.03	21.48	46.27	21.16	30.10	15.02
✓	✓		Ensemble	62.08	45.37	32.82	22.47	47.67	21.68	30.03	15.04
✓	✓		KMeans	62.43	45.78	32.90	22.19	47.61	21.87	30.76	15.24
✓	✓	✓	Ensemble	63.12	46.37	34.08	23.71	47.92	21.72	29.52	14.99
✓	✓	✓	KMeans	**65.09**	**48.93**	**36.02**	**24.78**	**49.13**	**22.79**	**33.41**	**15.72**



	Caption:	SPICE:
First:	A person is reading an open book with papers.	30.0
Second:	A woman sitting on her bed in an room.	31.6
Third:	A man and woman playing in an room with some chairs.	9.5
Ours:	A woman is sitting on the bed reading an open book.	40.0

図 1 単視点入力（UpDown）と三視点入力（KMeans，“Ours”）による生成例と SPICE による評価値．“First”・“Second”・“Third” はそれぞれ，一人称・二人称・三人称視点を表す．

3.　実験

3.1　データセット

ベースラインの UpDown モデルは，顕著領域検出器を Visual Genome データセット [6] で事前学習し，抽出された領域特徴を基にキャプション生成器を Microsoft COCO データセット [7] で学習した．Ensemble モデル・KMeans モデルでは，学習済み UpDown モデルを利用し，入力画像の変更・顕著領域クラスタリングの導入を行った．また，第四人称視点の画像組と説明文から構成される評価用データセットを新たに構築した．九州大学の保有する情報構造化空間の実験環境 Big-Sensor Box [3] において本研究の想定シナリオを再現し，第四人称視点に準拠する三種類の映像組と対応する参照説明文を収集した．なお，説明文のアノテーションは，記録映像ごとに 5 パターンの説明文を付与した．本実験では，一回の撮影で二人の被験者がウェアラブルカメラ（GoPro HERO）を装着する．一方が居住者役として，読書や炊事などの日常行動を実施する．もう一方はロボット役として，居住者を追尾しながら近くで監視し，必要に応じて干渉する．また，三人称視点として，別のカメラを両者が観測できる位置に固定した．

3.2　文章間類似度に基づく定量的評価

前章で述べたデータセットを用いて，単視点画像を UpDown，2 視点・3 視点画像を Ensemble あるいは KMeans に入力し，画像ごとの生成説明文と参照説明文の類似度評価値を比較する．評価指標として，画像キャプション生成分野で一般的な BLEU，ROUGE-L，METEOR，CIDEr-D，SPICE を採用する．表 1 に，評価結果をまとめる．全ての評価指標において，提案手法である三つの画像を入力した Kmeans モデルが最も高い評価値を示した．図 1 に，単視点を入力した UpDown と三視点を入力した KMeans の生成例を示す．

4.　まとめ

本稿では，情報構造化空間における一人称・二人称・三人称視点画像を入力とした説明文生成手法について述べた．また，新規構築したデータセットを用いて生成した説明文の評価実験を行い，提案手法の有効性を示した．

謝辞

本研究は JSPS 特別研究員奨励費 19J12159 の助成を受けたものである．

参　考　文　献

[1] M. Bolanos, M. Dimiccoli and P. Radeva: “Toward Storytelling from Visual Lifelogging: An Overview”, *IEEE Transactions on Human-Machine Systems*, vol.47, no.1, pp.77–90, 2016.

[2] C. Fan, Z. Zhang and D. J. Crandall: “DeepDiary: Lifelogging Image Captioning and Summarization”, *Journal of Visual Communication and Image Representation*, vol.55, pp.40–55, 2018.

[3] R. Kurazume, Y. Pyo, K. Nakashima, A. Kawamura and T. Tsuji: “Feasibility Study of IoRT Platform “Big Sensor Box””, *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA)*, pp.3664–3671, 2017.

[4] 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.

[5] 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”, *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp.6077–6086, 2017.

[6] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei: “Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations”, *Int. Journal of Computer Vision (IJCV)*, vol.123, no.1, pp.32–73, 2017

[7] 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.