物体の詳細情報を考慮した時空間シーングラフによる案内文生成

ER20046 鈴木颯斗　　指導教授：藤吉弘亘

1. はじめに

自動車に広く利用されるナビゲーションシステムは，GPS と地図データを基盤として周辺情報をルールベースに処理して案内を行う．そのため，複雑な周辺状況では運転者の誤解を招く可能性がある．一方で，人間が案内を行う場合，視界に含まれる動的な物体などの情報を用いて指示することが可能である．こうした人間のような案内を行うことでドライバーの直感的な理解を目指したアプローチが Human-like Guidance(HlG) である．本研究では，HlG の実現に向け，シーン画像から認識した物体をノードとしたシーングラフとして表現して文章生成モデルにより視界情報に基づいた案内文の生成を行う．

2. Transformer を用いた文章生成

Transformer は，Attention 機構を導入した言語モデルである．その柔軟性と拡張性は高く，言語とは異なる種類のデータに適応することができる．CPTR[1] は，画像入力に対応した Transformer ベースモデルであり，与えられた画像について説明する文章の生成が可能である．一方で，画像に多数の物体が含まれる場合，曖昧な文章が生成される可能性が高い．HlG のような特定の物体に注目した文章を生成するには画像の入力では不十分である．

3. 提案手法

本研究では，空間情報及び時系列情報を時空間シーングラフとして表現し，HlG に適した案内文の生成手法を提案する．本手法の概略図を図 1 に示す．

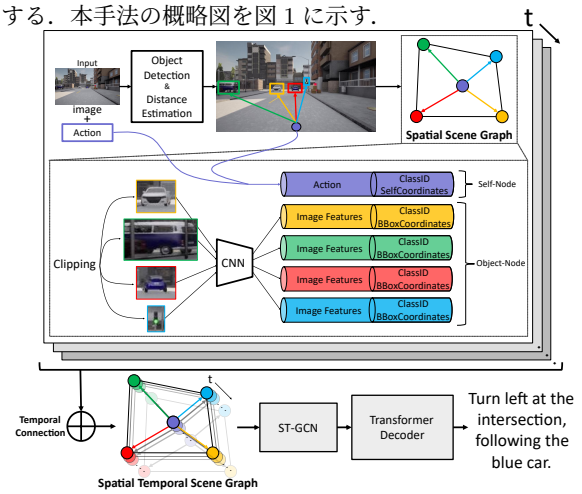

図 1: 提案手法

3.1. 時空間シーングラフの生成

始めに，YOLOv8 を用いて画像に含まれる物体（車両や信号機）の検出を行う．検出した物体をオブジェクトノードとしてグラフ化する．各ノードには，YOLOv8 が出力した ClassID と Bounding Box(BBox) 座標，抽出した BBox 領域内の特徴を用いる．また，自車両との関係性を考慮するために自車両を示すセルフノードを追加する．セルフノードには，ClassID，自車両座標とともに自車両の動作情報 “right”，“left”，“straight”）を用いる．オブジェクトノードとセルフノードを接続するエッジは，BBox サイズと実際の物体サイズを考慮して算出した物体間距離を重みとして与える．このシーングラフを各時刻で求めて時空間シーングラフとする．時系列接続を行うことで時間情報を考慮したグラフとなる．

3.2. 文章生成モデル

時空間シーングラフを ST-GCN[2] に入力し，物体の位置関係や時間的な変化を捉えた特徴量を抽出する．そして，ST-GCN で抽出した特徴量を Transformer-Decoder に入力し，案内文を生成する．学習時，Transformer-Decoder は入力シーンに対応した案内文と抽出したグラフ特徴量から単語列を逐次的に推測してデータセットに含まれる案内文の統計的な特徴を獲得する．推論時，グラフ特徴量と文章の開始を示すトークンが入力され，獲得した特徴を基に単語列を逐次的に推測することで案内文を生成する．

4. データセット

本研究の有効性を評価するために，CARLA Simulator を用いて HlG 用のデータセットを作成した．全 160 シーン，計 10,219 フレームで構成されており，交差点付近の走行シーンと案内文，交差点における動作情報が含まれる．案内文は，先頭に “Turn left” 等の動作情報，その後ろに適切な物体を中心とする指示を含む形式とする．

5. 評価実験

提案手法の有効性を検証する評価実験を行う．

5.1. ベースラインと比較条件の設定

シーン画像のみを用いた手法をベースラインとする．また，両方のアプローチにおいて時系列情報の有無による比較を行う．各モデルの詳細を表 1 に示す．

表 1: 提案手法とベースラインのモデル構造

	時系列	Encoder	Decoder
ベースライン		ResNet-18	Transformer
	✓	ResNet3D-18	Transformer
提案手法		GCN	Transformer
	✓	ST-GCN	Transformer

5.2. 実験設定

エポック数を 100，学習率を 0.0001 とし，時系列情報を含む場合は入力を 5 時刻分とする．また，評価指標として BLEU，METEOR，ROUGE を用いる．

5.3. 実験結果

定量的評価を行った結果を表 2 に示す．表 2 より，時系列情報を含む提案手法が最も高精度であることが確認できる．

表 2: 評価結果

	時系列	Bleu-4	METEOR	ROUGE
ベースライン		0.202	0.512	0.533
	✓	0.180	0.514	0.566
提案手法		0.209	0.575	0.598
	✓	**0.224**	**0.603**	**0.624**

案内文生成結果の例を図 2 に示す．図 2 より，提案手法において時系列情報を含めることで注目対象の進行方向を考慮した詳細な案内方法による文章が生成された．これは，時空間シーングラフにより注目対象とその動作の関係を捉えることが可能になったためと考えられる．

図 2: 各モデルの案内文の生成例

6. おわりに

本研究では，シーン画像から時空間シーングラフを生成する方法を提案し，文章生成モデルによる案内文生成と精度の評価について検証を行った．結果より，注目対象を中心とした適切な案内文の生成を実現した．今後は，案内方法の表現を上げるために正解案内文の増強を検討する．また，HlG に適した評価指標の検討を行う．

参考文献

[1] W. Liu, *et al.*, “CPTR: Full Transformer Network for Image Captioning”, CVPR, 2021.

[2] S. Yan, *et al.*, “Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition”, AAAI, 2018.