時空間シーングラフを用いた案内文生成の高精度化と視覚的説明に関する研究

TR24009 鈴木颯斗　　指導教授：藤吉 弘亘

1. はじめに

運転者の認知的負担を抑えつつ，走行状況を直感的に理解できる案内を提供することが，快適な運転ナビゲーション支援では重要である．このようなナビゲーションシステムの実現においては，車両が運転手に対して運転シーンに合わせて適切な情報を提示する必要がある．既存システムは，地図情報に基づいた経路案内や定型的な音声案内が主流であり，複雑かつ動的に変化する走行環境では，運転手が直感的に状況を把握することは容易ではない．このような課題に対し，周囲の状況を踏まえて人間のように案内を行う Human-like Guidance に関する研究が注目されている．

本研究では，Human-like Guidance の実現に向け，視界情報を基にした環境認識と自然言語生成を統合し，運転手が直感的に理解可能な案内文を生成することを目標とする．走行シーンにおける判断対象となる情報は多岐にわたり，特に時間的変化を伴う環境認識では，情報量の増加が生成精度や安定性に影響を及ぼす可能性がある．そこで本研究では，車両の視界情報から得られるオブジェクトの空間的・時間的関係を時空間シーングラフとして表現し，これを基に案内文を生成する手法を提案する．さらに，生成したシーングラフに対し，Graph Attention Networks(GAT)[1] を用いて案内に重要な対象を強調しながら情報を統合することを目指す．そして，推論時に得られる Attention を可視化することで，生成された案内文に対する視覚的説明を可能とする．

2. 関連研究

本研究では，車両の視界情報を基に環境を理解し，運転状況に即した案内文を生成することを目的としている．本章では，この目的に関連する技術を述べる．

2.1 動画像からのキャプション生成

キャプション生成は，画像または動画像を入力とし，その内容を自然言語による文章として生成するタスクである．本タスクでは，一般に視覚特徴を抽出するエンコーダと，抽出された特徴に基づいて文章を生成するデコーダからなる Encoder-Decoder 構成が採用される．視覚特徴抽出の手法として，画像を対象とする場合には CNN，動画像を対象とする場合には 3DCNN や時系列情報を考慮可能な Transformer ベースのモデルが広く用いられる．文章生成には，Transformer に代表される自己回帰型の言語モデルが用いられ，Cross-Attention 機構を通じて視覚特徴と単語列を統合しながら逐次的に自然言語の説明を生成する．これにより，入力画像や動画像の内容と意味的に整合性のあるキャプション生成が可能となる．

2.2 シーングラフによる環境理解

Graph Neural Network を視覚認識へ応用した手法として，画像中のオブジェクトをノードとして表現し，その関係性をグラフ構造として扱うシーングラフに基づく手法が提案されている．Graph R-CNN[2] は，物体検出モデルによって得られたオブジェクト間の関係をシーングラフとして明示的に構築することで，画像の構造的理解が向上することを示している．

3. 提案手法

本研究では，走行シーンの動画像から得られるオブジェクトの時空間的関係を時空間シーングラフとして構築し，GAT を用いた Graph-to-Text モデルにより案内文を生成する手法を提案する．また，Graph Encoder における Attention を可視化することで，案内文生成の判断根拠を視覚的に提示し，モデルの解釈性と信頼性の向上を図る．提案手法の全体構成を図 1 に示す．

3.1 マルチオブジェクトトラッキング

交通シーンには多様なオブジェクトが存在し，その外観も大きく変化する．従来のシーングラフ構築では，オブジェクトをノードとして定義した場合に，この多様な外観情報を含めることが困難である．そこで本研究では，Open-Vocabulary 物体検出モデルである YOLO-World[3] を用いる．YOLO-World は画像特徴とテキスト特徴を統合したクロスモーダル表現により，未学習クラスの zero-shot 検出が可能である．本手法では，テキストで指定したクラスラベルを直接ノードのラベルとして利用することで，ノード表現を簡潔かつ解釈可能な形式に統一する．

また，オブジェクト自身の時間的な差分をシーングラフとして表現する場合，検出オブジェクトをフレーム間で一貫して追跡する必要がある．本手法では追跡アルゴリズムである BoT-SORT[4] を用い，各オブジェクトに一貫した ID を付与する．これにより，シーングラフ構築時に時間方向の解析が可能となる．

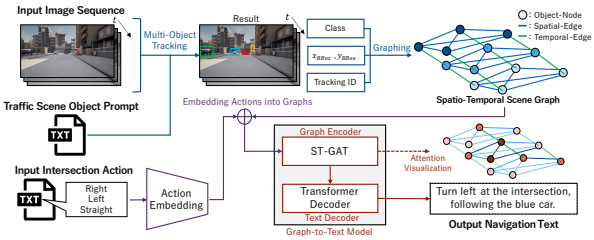

図 1: 提案手法のアーキテクチャ

3.2 時空間シーングラフの構築

動画像が与えられたとき，前述したマルチオブジェクトを行い，検出オブジェクトの位置・クラス情報・追跡情報を用いて時空間シーングラフ G を式 (1) として構築する．

$$G = \{\boldsymbol{V}, \boldsymbol{E}\} \tag{1}$$

ここで，$\boldsymbol{V}$ はノード集合，$\boldsymbol{E}$ はエッジ集合である．

ノード集合 $\boldsymbol{V}$ の定義： 各フレーム t におけるノード集合 $\boldsymbol{V}_t$ を，マルチオブジェクトトラッキングによって得られた検出オブジェクトの集合として，式 (2) のように定義する．

$$\boldsymbol{V}_t = \{v_i^t \mid (B_i^t, c_i^t, id_i^t) \in \mathcal{D}_t\} \tag{2}$$

ここで，$\mathcal{D}_t$ はマルチオブジェクトトラッキングの出力を表し，B はオブジェクト毎の境界ボックス座標，c はクラスラベル，id は ID 化されたトラッキング情報を示す．

エッジ集合 $\boldsymbol{E}$ の定義： 同一フレーム内のオブジェクト間の関係性を空間的エッジ $\boldsymbol{E}_{\text{spatial}}$ として式 (3) のように定義する．

$$\boldsymbol{E}_{\text{spatial}} = \{((v_i^t, v_j^t), w_{ij}^t)\}, w_{ij}^t = \| \tilde{B}_i^t - \tilde{B}_j^t \|_2 \tag{3}$$

ここで，w_{ij}^t はエッジに対する重みを示し，オブジェクト間のユークリッド距離を付与する．これにより，フレーム毎のオブジェクト間の動的変化をグラフ内に表現する．

続いて，時系列方向におけるオブジェクト間の関係性として時系列エッジ $\boldsymbol{E}_{\text{temporal}}$ を式 (4) のように定義する．

$$\boldsymbol{E}_{\text{temporal}} = \{((v_i^t, v_j^{t+1})) \mid id_i^t = id_j^{t+1}\} \tag{4}$$

この処理では，前後のフレームでトラッキング ID が一致するノードに対してエッジが接続される．

最終的に，エッジ集合 $\boldsymbol{E}$ は式 (5) となる．

$$\boldsymbol{E} = \boldsymbol{E}_{\text{spatial}} \cup \boldsymbol{E}_{\text{temporal}} \tag{5}$$

3.3 シーングラフへの Action 埋め込み操作

案内文生成では，シーンの状況だけでなく，自車の行動 Action（右折・直進など）を考慮することが重要である．本手法では，ナビゲーション時に与えられる Action をテキスト形式として入力し，埋め込み層を通して Action 特徴を得る．そして，得られた Action 特徴を前段で生成されたシーングラフ内のすべてのノード特徴へ統合する．事前にシーングラフに Action を埋め込むことで，Action に基づいて着目すべきノードが強調されるような効果を図る．

表 1: 各モデルで生成された案内文の精度結果

Method	5 frame				10 frame				15 frame			
	B-1	B-4	M	R	B-1	B-4	M	R	B-1	B-4	M	R
3DCNN	0.568	0.322	0.575	0.643	0.551	0.292	0.538	0.617	0.519	0.268	0.515	0.601
3DResNet	0.459	0.197	0.446	0.559	0.448	0.173	0.439	0.547	0.457	0.180	0.449	0.534
VTN	0.583	0.337	0.578	0.565	0.412	0.142	0.378	0.537	0.379	0.099	0.377	0.471
ViViT	0.524	0.266	0.538	0.592	0.540	0.285	0.551	0.603	0.549	0.274	0.559	0.611
Ours	**0.610**	**0.363**	**0.635**	**0.668**	**0.617**	**0.382**	**0.646**	**0.675**	**0.631**	**0.388**	**0.649**	**0.677**

※ B-1：BLEU-1, B-4：BLEU-4, M：METEOR, R：ROUGE

3.4 Graph-to-Text モデル

本研究では，時空間シーングラフから文章の生成を行う Graph-to-Text モデルを提案する．Graph-to-Text モデルは，グラフの特徴抽出を行う Graph Encoder と文章生成を行う Text Decoder で構成される．Graph Encoder では，空間方向と時系列方向に分けて Attention を適用し，特徴抽出を行う Spatial Temporal GAT (ST-GAT) を構築する．Text Decoder では，Graph Encoder により抽出されたグラフ特徴量から，Transformer Decoder を用いて文章の生成を行う．また，モデルの推論時，最終層における各エッジに対する Attention スコアをグラフ上に可視化することで，モデルの判断根拠の解釈を可能とする．

4. データセット

ナビゲーションタスクには，走行車両の車載カメラ映像と対応する案内文のペアからなるデータセットが必要となる．本研究では案内文生成に特化したデータセットを CARLA Simulator を用いて独自に作成する．撮影には 8 つのマップを用い，撮影条件は以下の通り設定する．

- フレームレート：10 fps
- 天候条件：ClearNoon, WetNoon
- 撮影範囲：交差点約 50m 手前から交差点通過直後

案内文のアノテーションは手動で実施し，注目対象に基づいた案内文を作成する．作成したデータセットは，合計 160 シーン，計 10,219 フレームで構成される．各シーンには，前述した案内文，進行方向における動作情報が含まれる．

5. 評価実験

評価実験を通じて提案手法の有効性を検証する．本実験では，ベースライン手法との比較，入力フレーム数が 5 フレーム，10 フレーム，15 フレームにおける異なるフレーム長が案内文の生成精度に与える影響について分析する．評価には，BLEU，METEOR，ROUGE を用いる．

5.1 ベースライン手法

ベースライン手法として動画像から直接特徴量を抽出する手法を用いる．具体的には，提案手法におけるシーングラフを構築する過程と Graph Encoder を Video Encoder に置き換える．本実験では，CNN および Transformer をベースとした，3DCNN，3DResNet，Video Transformer Network (VTN)，Video Vision Transformer (ViViT) を用いる．

5.2 実験条件

学習設定は，学習率 1.0×1.0^{-4}，エポック数 100，バッチサイズ 32，Dropout 率 0.3 とする．学習の最適化アルゴリズムには AdamW を用いる．これらの設定は，提案手法およびベースライン手法の全てのモデルで統一する．

5.3 定量的評価

提案手法およびベースライン手法の各モデルで生成された案内文の精度について定量的評価によって比較を行う．評価結果を表 1 に示す．結果より，提案手法は全てのフレーム数において他の手法を上回る精度を達成しており，フレーム数が増加するほどより顕著に精度が向上していることが確認できる．

5.4 定性的評価

提案手法およびベースライン手法の各モデルで生成された案内文について定性的に評価する．各手法における案内文生成結果の例を図 2 に示す．結果より，Groud Truth と同様の “yellow car” を中心とした案内文を生成できているものは提案手法のみであり，最も適切な説明となっている．ベースライン手法においては，最も動作の変化が大きい “black car” もしくは画像内に存在しないオブジェクトを注目しており，不適切な説明となっている．

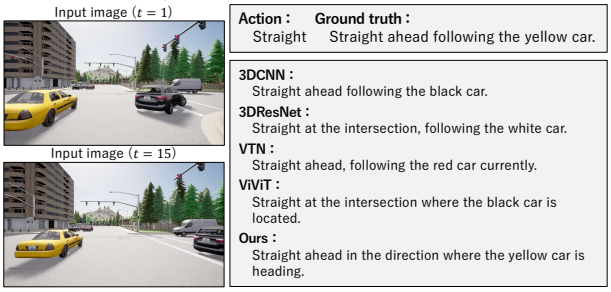

図 2: 各手法における案内文生成結果

次に，提案手法における案内文生成において，推論時の Attention を時空間シーングラフ上に可視化する．Attention の可視化結果を図 3 に示す．結果より，グラフ上では “black car” に着目しており，生成案内文の着目しているオブジェクトと一致する．また，エッジはオブジェクト間の関連度として解釈することができる．したがって，モデルが生成した案内文の判断根拠をグラフを通して視覚的に説明可能であることを示している．

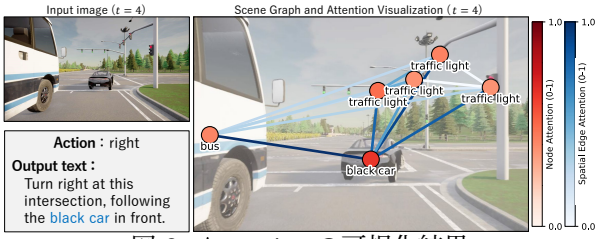

図 3: Attention の可視化結果

6. おわりに

本研究では，車両の視界情報から動的な環境を理解し，運転手に直感的な案内文を生成する手法を提案した．走行シーンのオブジェクト関係を時空間シーングラフとして表現し，Graph-to-Text モデルにより案内文生成を行った．また，GAT による重要情報の強調と Attention 可視化による判断根拠の提示を実現した．評価実験では，提案手法が CNN や Transformer ベースの Video Encoder を用いたベースライン手法よりも高い精度を示し，特に長期間の情報統合において有効性が確認された．さらに，定性的評価および Attention 可視化から，モデルが適切な対象に注目し案内文を生成していることを確認した．

今後の課題として，より複雑な環境や多様な運転シナリオへの適用とその有効性の検証が挙げられる．

参考文献

[1] V.Peter, *et al.*, “Graph Attention Networks”, ICLR, 2018.

[2] Y.Jianwei, *et al.*, “Graph R-CNN for Scene Graph Generation”, ECCV, 2018.

[3] T.Cheng, *et al.*, “YOLO-World: Real-Time Open-Vocabulary Object Detection”, CVPR, 2024.

[4] N.Aharon, *et al.*, “BoT-SORT: Robust Associations Multi-Pedestrian Tracking”, arXiv, 2022.

研究業績

[1] H. Suzuki, *et al.*,“Enhancing Navigation Text Generation and Visual Explanation Using Spatio-Temporal Scene Graphs with Graph Attention Networks”, ITSC, 2025.

(他 学会発表 3 件)