

複数歩行者間の関係性を考慮した 注視対象物推定の全体最適化の検討

辰己 弘征^{1,a)} 出口 大輔^{1,b)} 伊藤 誠悟^{1,c)} 村瀬 洋^{1,d)}

概要

本発表では、複数の歩行者が往来する環境における自律移動ロボットの安全かつ円滑な移動を目的として、複数の歩行者に対して注視対象物を推定する手法を提案する。ロボットが歩行者との接触を防ぎつつ円滑な移動をするためには、周囲の複数歩行者間のインタラクション理解が重要であり、各歩行者の注視対象物はそのようなインタラクション理解のために必要な情報である。従来の注視対象物推定 [1] では、画像中の個別の歩行者に注視対象物推定を行っていたが、各歩行者の注視対象物には何らかの関係性があると考えられる。本発表では、注視対象物推定の精度向上を目的として、複数歩行者間の関係性を考慮した注視対象物推定の全体最適化手法を提案する。

1. はじめに

近年、ショッピングモール等への自律移動ロボットの導入が進んでいる。このような環境におけるロボットは、歩行者間のインタラクションを理解し、円滑な走行をすることが求められる。図 1 のような環境では、歩行者 A と歩行者 B は互いを見合っているため同一グループだと推測できる。そのため、ロボットがグループの間を割って入ることは適切ではない。また、歩行者 C は反対側から近づく歩行者 B を注視しているため、歩行者 C は歩行者 B を避けようとしていると推測できる。そのためロボットは、歩行者 C に注意しながら適切な経路を選択する必要がある。このように周囲の複数歩行者間のインタラクションを理解することで、ロボットは安全かつ歩行者に配慮した経路選択が可能になる。そしてこの例からわかるように、歩行者の注視対象物はインタラクション理解に必要な不可欠な要素である。また歩行者間のインタラクションの中では、歩行者 D のように明確な注視対象物を持たない場面も多数存在する。明確な注視対象物を持たない歩行者に対して特定

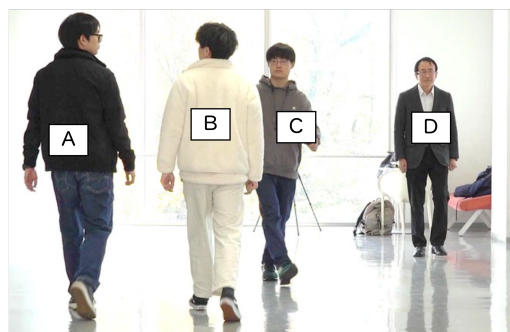


図 1 複数歩行者間のインタラクションの例

の注視対象物を持つと推論すると、インタラクション理解を誤る可能性がある。そのため、注視対象物の有無を含めた注視対象物推定が必要である。また、歩行者の注視対象物は他の歩行者からの影響を受けることがわかっている。Gallup ら [2] の研究では、歩行者が他の歩行者の視線方向を模倣する現象が実験的に確認されている。そこで本発表では、歩行者同士のインタラクションの中で、注視対象物の有無を含む注視対象物推定タスクに取り組む。具体的には、精度向上を目的として複数歩行者の関係性を考慮した注視対象物推定の全体最適化に取り組む。これまでに歩行者間のインタラクションを含むデータセットがいくつか公開されているが [5]、これらは追跡や物体検出を目的としており、注視対象物のアノテーションを含まない。また、歩行者は明確な注視対象物を持たない場合が多いため、注視対象物推定のための分類器の学習が難しいという問題がある。そこで我々は注視対象物推定のために、「インタラクション理解に向けた気づき認識データセット」（気づき認識データセット）[3] を構築した。本データセットでは予め作成した 20 種類のシーンに従って、複数人の歩行者間のインタラクションを撮影し、作成したシーン通りにアノテーションを行った。本発表では、気づき認識データセットを利用した、注視対象物推定ネットワークを提案する。

2. 関連研究

2.1 頭部の特徴に依存した注視対象物推定

Tonini ら [4] は、第三者視点の画像を用いて人々の注視対

¹ 名古屋大学大学院 情報学研究科

a) tatsumih@vislab.is.i.nagoya-u.ac.jp

b) ddeguchi@nagoya-u.jp

c) iseigo@vislab.is.i.nagoya-u.ac.jp

d) murase@nagoya-u.jp

象物を推定する手法として、Gaze Object Transformer を提案している。視線方向のアノテーションを含む GazeFollow データセット [6] に対して、注視対象物のアノテーションを追加したものを実験に利用している。具体的には、画像から頭部検出を行い、抽出した頭部特徴と周辺の物体から抽出した特徴をマッチングすることで注視対象物を予測する。そのため、注視対象物推定の精度は、頭部特徴に左右される。歩行者が自由に行き交うショッピングモール等の環境では、ロボットから見た歩行者が遠方になる可能性があるため、信頼性の高い頭部特徴の抽出が困難になる可能性がある。

2.2 歩行者の注視対象物推定

ショッピングモール等における自律移動ロボットを考えた場合、単眼カメラと歩行者の距離が遠く頭部が不鮮明に撮影されるため、従来の頭部画像（主に眼球領域）に依存した注視対象物推定手法の適用が困難である。そこで我々の研究グループでは、頭部画像のみに依存しない、歩行者全体の特徴を用いた注視対象物推定に取り組んでいる。畑ら [8] の研究では、屋外の交通シーン画像から歩行者の骨格や位置座標を取り出し、歩行者が自車両を見ているかを推定している。村上ら [1] の研究では、屋外の交通シーン画像から、歩行者の注視対象物（注視対象物の候補は歩行者と物体）を推定する PEGOTransformer を提案している。

一方、自律移動ロボットを取り巻く歩行者間のインタラクションでは、歩行者が明確な注視対象物を持たない場合も数多く存在する。そのため、注視対象物の有無を含めた注視対象物推定が必要である。しかし PEGOTransformer では、歩行者が注視対象物を持つことを前提として注視対象物を推定しているため、注視対象物の有無を推定できない。

2.3 気づき認識データセットの構築

村上らは、注視対象物推定向けの「歩行者の注視対象物データセット」[7] を構築した。これは一般公開されている交通シーンデータセットに対して、歩行者の注視対象物をアノテーションしたものである。しかしこれは屋外の交通シーンを対象としており、本発表で想定する自律移動ロボットを取り巻く環境とは異なる。また、一般公開された交通シーンデータセットに対して注視対象物をアノテーションしたものであるため、注視対象物の真値が不明である。一方で、ショッピングモール等の自律移動ロボットを取り巻く環境を想定したデータセットはいくつか存在するものの、注視対象物がアノテーションされていない。そこで我々は、気づき認識データセットを構築した。具体的には自律移動ロボットを取り巻く環境を想定して、歩行者同士のインタラクションを含む複数のシーン（注視対象物となる歩行者と歩く方向を指定）をあらかじめ用意し、シー

ンに沿って撮影とアノテーションを行った。各歩行者に対するアノテーションには、歩行者の ID、注視対象物の ID、明確な注視対象物を持たないラベル、BBox の座標が含まれる。

3. 歩行者の注視対象物推定

本節では、注視対象物推定ネットワークについて述べる。注視対象物推定ネットワークは、PEGOTransformer [1] を改良したものであり、複数歩行者間の関係性を考慮した注視対象物推定の全体最適化を行う。注視対象物推定ネットワークは、画像から画像特徴量を獲得する Backbone、各物体に対応する特徴を獲得する Deformable Transformer、各物体に対応する特徴を同時に処理し、物体同士の関係を捉える Projection Layer (Transformer Encoder)、各歩行者に対応する特徴と、歩行者の注視対象物候補に対応する特徴の関係を捉え、注視対象物の推定結果を出力する Projection Layer (Cross Attention)、Label Generator の 5 つのモジュールから構成される。以下で各モジュールについて紹介する。

3.1 注視対象物推定ネットワークの構成

1) **Backbone** 入力画像 $\mathbf{x} \in \mathbb{R}^{C \times H \times W}$ を ResNet [10] に入力して画像特徴量を獲得する。

2) **Deformable Transformer** Deformable Transformer [9] は、Encoder と Decoder から構成される。Backbone から得られた画像特徴量を入力とし、各物体に対応する Object Query ($\mathbf{OQ} \in \mathbb{R}^C$)、BBox 座標、Class を出力する。

Deformable Transformer Encoder: CNN バックボーンから得られた特徴マップ $\mathbf{f} \in \mathbb{R}^{C \times H \times W}$ を平坦化し、位置エンコーディングを付加して Encoder 層に入力する。Self-attention により画像全体の空間的関係性を学習した特徴表現を生成する。

Deformable Transformer Decoder: 学習可能な N 個の Object Query $\mathbf{OQ} = \{\mathbf{oq}_1, \mathbf{oq}_2, \dots, \mathbf{oq}_N\}$ をランダムに初期化して入力する。Self-attention で Object Query 間の関係を学習し、Cross-attention で Encoder 出力と各 Object Query を対応付ける。これにより各 $\mathbf{oq}_i$ は画像中の特定物体に対応した特徴量となり、FFN を通して BBox 座標と Class が予測される。

3) **Projection Layer (Transformer Encoder)** Projection Layer (Transformer Encoder) [1] は、Deformable-Transformer から得られた Object Query $\mathbf{OQ}$ 、BBox 座標、Class、そして事前に決定した Entity type (注視対象物推定する歩行者を Entity_type = 0, 注視対象物推定しない歩行者を Entity_type = 1, 無視する物体を Entity_type = 2) について、これらのエンベディングを入力とし、Transformer Encoder [11] により各種エンベディングの関係性、物体同士の関係性を学習して、特徴表現を出力する。村上らの

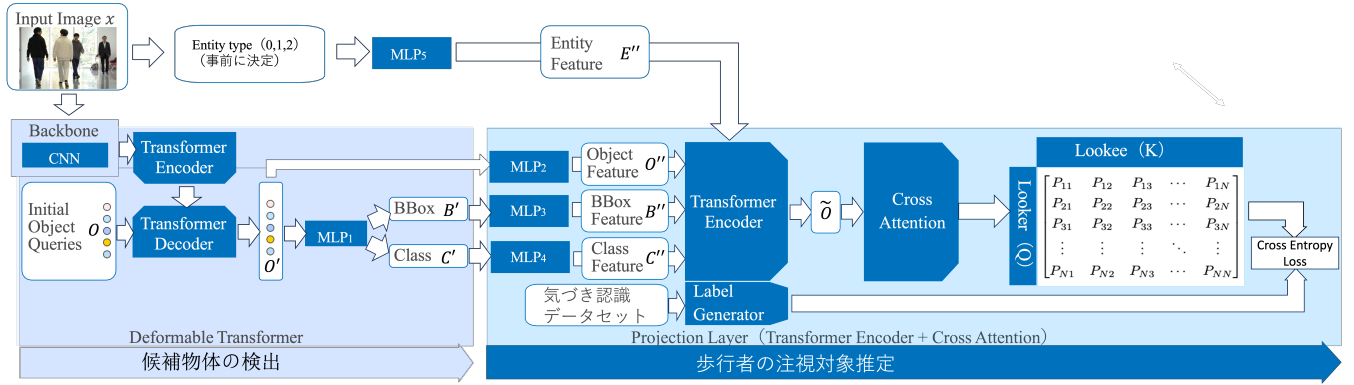


図 2 注視対象物推定ネットワーク

Projection Layer では、1 サンプルが歩行者と注視対象物候補のペアとして扱われ、各歩行者の注視対象物推定が独立して行われていた。そのため、同一画像内の他の歩行者の注視対象物の情報や、複数人が注目する物体の特徴といった画像全体のコンテキスト情報が活用できない制約があった。本手法では、Transformer Encoder 内で画像内の全歩行者の Object Query を含む各種エンベディングを同時に処理することで、画像内の複数歩行者の関係性を考慮した注視対象物推定の全体最適化を行う。具体的には、Transformer Encoder の self-attention により、同じ対象を見る歩行者同士の Object Query が類似した特徴を持つよう学習する。

4) Projection Layer (Cross Attention) Projection Layer (Cross Attention) は、Projection Layer (Transformer Encoder) の特徴表現のうち Object Query ($OQ \in \mathbb{R}^C$) に由来する特徴のみを入力として、全歩行者の注視確率を出力する。注視対象物推定する歩行者 (Entity_type = 0, looker) の特徴を Query (Q) とし、注視対象物候補 (Entity_type = 0.1, looker) の特徴を Key (K) および Value (V) とする。各歩行者について、 $Q \times K^T$ の内積計算によりアテンションスコアを算出し、softmax 関数により注視確率分布を得る。注視対象物が存在する場合は対応する物体への注視確率が高くなり、何も見ていない場合は自分自身への注視確率が高くなるよう学習する。

5) Label Generator Label Generator は学習に用いる教師ラベルである PEGO 行列 $P \in [0, 1]^{N \times N}$ [1] を生成する。PEGO 行列は画像中の物体間の注視関係を表し、 i 番目の Object Query oq_i に対応する歩行者が j 番目の Object Query oq_j に対応する物体を注視している場合、 $P_{i,j} = 1$ とする。また、 i 番目の Object Query oq_i に対応する歩行者が明確に注視対象物をもたない場合、 $P_{i,i} = 1$ とする。

3.2 損失関数

注視対象物推定をする際に「特定の注視対象物を持つ」場合と「明確な注視対象物を持たない」場合の間に生じる

クラス不均衡によって、推論に偏りが生まれる可能性がある。このクラス不均衡問題に対処するため、重み付きクロスエントロピー損失を用いる。加えて、画像内の複数歩行者の注視対象物を最適化することを目的として、各画像 (サンプル) 内で歩行者の損失を平均化してからバッチ内の画像間で平均を取る段階的平均化手法 (画像毎平均損失) を導入する。これにより、同一画像内の歩行者間の注視対象物の関連性を保持しながら学習を行う。

$$L = \frac{1}{B} \sum_{i=1}^B \left[\frac{1}{N_{lookers_i}} \sum_{j \in lookers_i} w_{i,j} (-\log(\hat{P}_{i,j})) \right] \quad (1)$$

ここで、 B はバッチサイズ、 $lookers_i$ は画像 i 内の注視対象物推定される歩行者の集合、 $N_{lookers_i}$ は画像 i 内の注視対象物推定される歩行者数、 $w_{i,j}$ は正解ラベルの種類に応じた重み係数である (j は注視対象物の正解インデックスを示す)。重み係数は以下のように定義される

$$w_{i,j} = \begin{cases} w_{no-gaze} & \text{if } j^* = i \\ w_{gaze} & \text{if } j^* \neq i \end{cases} \quad (2)$$

$w_{no-gaze}$ は正解が「明確な注視対象物を持たない」場合の補正重み、 w_{gaze} は正解が「特定の注視対象物を持つ」場合の補正重みであり、候補数に応じた動的調整により安易な「明確な注視対象物を持たない」の選択を抑制する。

4. 実験

注視対象物推定ネットワークを用いて、歩行者の注視対象物推定を行う。実験設定と結果について述べる。

4.1 実験設定

我々が構築した気付き認識データセットのうち、今回の実験で利用するものを表 1 に示す。表 1 の各値は、気付き認識データセットのうち、Deformable DETR [9] で検出したものである。学習時には Projection Layer のパラメータのみを更新対象とし、Backbone と Deformable DETR で訓練済みの重みを用いて初期化した。実験では、 $w_{i,j} = 1$ としている。比較対象として、画像ごとの平均化を行わず、



図 3 注視対象物推定の正解例（左）と一部失敗例（右）

バッチ内の全歩行者の損失を直接平均したものを損失（全歩行者平均損失）として返す実験も行う。

表 1 実験に利用したデータセットの内訳

項目	学習	テスト
画像枚数	13170	7669
特定の注視対象物を持つ歩行者数	15725	10625
明確な注視対象物を持たない歩行者	12779	7996
平均注視対象物候補数	2.5	3.0

4.2 結果

表 2 に推定の精度を示す。評価指標として Top-K accuracy を利用した。Top-K accuracy は注視対象物推定の結果で最も尤度が高いものから順に並べた時に上位 K 個以内に正解が入る割合を示す。図 3（左）は 4 人の注視対象物推定が全て正解している例であり図 3（右）は 4 人のうち 1 人の注視対象物推定が誤っている例である。図の青の矢印は特定の歩行者を注視していると推定して正解したことを示している。また、赤のバツ印は「明確な注視対象物を持たない」と推定して誤ったことを示し、青のバツ印は「明確な注視対象物を持たない」と推定して正解したことを示している。この例から、複数の歩行者間のインタラクションにおける注視対象物を捉えられている一方、複数人が同じものを見ているという状態を推定することができていないことがわかる。

表 2 2 つの損失関数を用いた注視対象物推定の結果

	Top1	Top2	Top3
画像毎平均損失	0.664	0.896	0.967
全歩行者平均損失	0.65	0.881	0.952
チャンスレート	0.25	0.50	0.75

5. 考察

実験では、画像内の複数歩行者の注視対象物を最適化する損失関数（画像毎平均損失）と、比較対象の画像ごとの平均化を行わずにバッチ内の全歩行者の損失を直接平均する損失関数（全歩行者平均損失）を用いて実験を行った。どちらの損失関数を用いた場合でも Projection Layer (Transformer Encoder) 内で画像内の全歩行者の各種エンベディングを同時に処理しているため、複数歩行者の関係

性を考慮した注視対象物推定を行っているが、損失関数によって、より明確に画像単位 of 複数歩行者の注視対象物推定の最適化を行った。

その結果 Top1-accuracy, Top2-accuracy, Top3-accuracy の結果において、画像毎平均損失を用いた場合、全歩行者平均損失を用いた場合に比べてわずかに向上した。しかし、図 1 のように、画像毎平均損失を用いた場合でも、同じ対象を見ている歩行者同士の関係を正確に学習できていないという課題がある。今後、これらの歩行者の関係を捉えることで、注視対象物推定の精度が向上することが期待される。また今回は、クラス不均衡が存在するにも関わらず、式 (2) の $w_{i,j}$ 値を常に 1 にして実験を行った。今後適切な重みを動的に作成することを検討する。

6. むすび

本発表では、PEGOTransformer [1] を改良し、歩行者の注視対象物を推定するネットワークを提案した。また、損失関数によってより明確に画像単位 of 歩行者の注視対象物を最適化することで注視対象物推定の精度が向上することを示した。今後の課題として、同じ対象を見ている歩行者同士の関係をより正確に学習できる注視対象物推定ネットワークの提案などが挙げられる。

謝辞

本研究は、トヨタ自動車株式会社および科学研究費補助金 23K28164 の支援によって行われた。本研究は著者ら独自の意見及び結論を反映したものである。

参考文献

- [1] Hiroto Murakami, Jialei Chen, Daisuke Deguchi, Takatsugu Hirayama, Yasutomo Kawanishi, and Hiroshi Murase, "Pedestrian's Gaze Object Detection in Traffic Scene," Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP), pp. 333-340, 2024.
- [2] Andrew C. Gallup, Joseph J. Hale, David J. T. Sumpter, Simon Garnier, Alex Kacelnik, John R. Krebs, and Iain D. Couzin, "Visual attention and the acquisition of information in human crowds", Proceedings of the National Academy of Sciences (PNAS), pp. 7245-7250, 2012.
- [3] 辰己弘征, 出口大輔, 村瀬洋, 伊藤誠悟, "複数歩行者間のインタラクション理解に向けた気づき認識データセットの構築," 電子情報通信学会論文誌, pp. 25-26, no.406, vol.124, MVE2024-34, 2025.
- [4] Francesco Tonini, Nicola Dall'Asen, Cigdem Beyan, Elisa Ricci, "Object-aware Gaze Target Detection," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 21860-21869, 2023.
- [5] Patrick Dendorfer, Hamid Rezaatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé, "MOT20: A benchmark for multi object tracking in crowded scenes," arXiv preprint arXiv:2003.09003, 2020.
- [6] Adrià Recasens, Aditya Khosla, Carl Vondrick, Antonio

- Torrallba, "Where are they looking?," Advances in Neural Information Processing Systems (NIPS), pp. 199-207, 2015.
- [7] 村上大斗, 出口大輔, 平山高嗣, 川西康友, 村瀬洋, 歩行者の注視対象データセットの構築, 情報処理学会研究報告, Vol.2023-CVIM-233 No.57, 2023
- [8] Ryusei Hata, Daisuke Deguchi, Takatsugu Hirayama, Yasutomo Kawanishi, Hiroshi Murase, "Detection of distant eye-contact using spatio-temporal pedestrian skeletons," IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), pp. 2730-2737, 2022.
- [9] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, Jifeng Dai, "Deformable DETR: Deformable Transformers for End-to-End Object Detection," Proceedings of the 9th International Conference on Learning Representations (ICLR), 2021.
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Deep Residual Learning for Image Recognition," Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, 2016.
- [11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin, "Attention is All you Need," Advances in Neural Information Processing Systems (NIPS), pp. 6000-6010, 2017.