交通シーンにおける歩行者の注視対象物推定の検討

村上 大斗[1,a)]　陳 嘉雷[1]　出口 大輔[1]　平山 高嗣[2,1]　川西 康友[3,1]　村瀬 洋[1]

概要

本発表では，交通シーン中における歩行者が注視している対象物（PEdestrian Gaze Object: PEGO）を推定するという新しいタスクと，Transformer を利用して PEGO を推定する手法を提案する．歩行者の行動予測において PEGO の推定は重要であり，より洗練された自動運転の実現にとって重要な情報である．しかしながら，自車両から離れた歩行者の解像度は低く，視線計測に基づく PEGO の推定は不可能である．そこで本発表では，歩行者の全身の特徴を利用したシンプルなアルゴリズムである PEGO Transformer を提案する．独自に構築した PEGO データセットを用いた評価により，PEGO Transformer の有効性を確認した．

1. はじめに

歩行者の注視対象物（PEdestrian Gaze Object: PEGO）推定は，自動運転車両等が歩行者の行動を予測するために解決すべき重要な技術課題である．例えば，図 1 のようにバイクを注視している歩行者は，車道に飛び出さず，少なくともバイクが通過するまで待つだろうと想定できる．このように，PEGO の推定結果は，歩行者が今後どのような行動をとるかを明らかにする重要な手がかりとなる．

これまでに，人間の注視対象を推定しようとする試みがいくつかなされている [7, 10]．しかし，これらは日常生活や小売店シーンを対象としていることから，カメラ画像に写る人物の割合が大きく，また人物と対象物の距離が非常に近いため，交通シーンとは対象物の密度や距離等の状況が大きく異なる．

そこで，これまでに我々は歩行者の注視対象物データセット [12] を構築した．本データセットでは，車載カメラ画像に写る歩行者が，同じ画像中に写るどの物体を注視しているかをアノテーションした．

本発表では，この歩行者の注視対象物データセットを用いた，交通シーンにおける歩行者の注視対象物 (PEGO) を推定する PEGO Transformer を提案する．PEGO Transformer は，入力画像から特徴を抽出する Feature Extractor，物体に対応する特徴を捉える Deformable Transformer Decoder，特徴を利用して PEGO の予測結果を生成する Projection Layer からなり，PEGO Transformer の学習に利用する教師ラベルを生成する Label Generator と合わせて 4 モジュールから構成される．

図 1: バイクを注視する歩行者

2. 関連研究

2.1 交通シーン以外での注視対象推定

Recasens ら [7] は，人間の視線推定のためのデータセットと手法である GazeFollow を提案している．この先駆的な研究は，人の行動予測における注視対象推定タスクの重要性を示している．このデータセットは視線方向の推定を目的としており，我々が目標とする注視する物体の推定ではない．視線方向の推定のみの場合，たとえ視線上で最近傍の物体を選択しても，二次元画像上では物体同士の奥行き関係が表現できないため，正しい注視対象物が得られるとは限らない．よって，従来の視線を推定対象とする手法ではなく，物体を推定対象とする，物体レベルの推定手法が必要である．

Wang ら [10] は，顧客がどの商品を注視しているかを物体レベルで推定する GaTector を提案している．この手法では対象人物の頭部特徴から注視領域を推定し，推定された注視領域と重なる物体を注視物体と判定している．しかし，GaTector の対象は小売店のように人物と対象物体が近いシーンであり，交通シーンとは人物と物体の位置関係や物体が配置される密度が大きく異なる．特に，車載カメラで撮影された歩行者の多くは距離が遠く，画像解像度が低くなる．そのため，小売店シーンで撮影された歩行

1 名古屋大学 大学院情報学研究科
2 人間環境大学 環境科学部
3 理化学研究所情報統合本部 ガーディアンロボットプロジェクト
a) hiroto.murakami@nagoya-u.jp

図 2: 歩行者注視対象物データセット収録例：対象歩行者を黄色の BBox，注視対象物を赤色の BBox，アノテーションした注視点を赤点で示す．

者よりも小さく，不鮮明に写る．よって，交通シーンでは GaTector が注視領域の推定に用いる頭部特徴の抽出が困難であり，PEGO 推定のためには新たな手法が必要である．

2.2 歩行者の注視対象推定

車載カメラ画像に写る歩行者は撮影距離やカメラ画角の関係から，小さく不鮮明に写ることが多いため，鮮明な頭部位置画像を必要とする既存の注視対象推定手法が適用できない．そこで我々の研究グループでは，頭部位置に依存せず，歩行者の全体的な外見に基づく注視対象推定に取り組んでいる．畑ら [3] の研究では，歩行者の骨格情報を用いることで歩行者が自車両を見ているかどうか（アイコンタクト）を推定している．しかし，自車両を見ていない歩行者が何を注視しているかは推定できない．

2.3 歩行者注視対象物データセットの構築

従来の注視対象推定タスク向けのデータセット [7, 10] は，推定対象が物体ではなく視線であり，人物と物体との位置関係が交通シーンとは大きく異なっている．そのため，交通シーンにおける PEGO 推定に利用することはできない．一方，交通シーンを対象とした物体検出等のデータセット [1, 7, 8] がいくつか存在するものの，いずれも歩行者の注視対象はアノテーションされていない．

そこで我々は，既存の交通環境データセットを拡張し，歩行者の注視対象物データセット [12] を構築した．このデータセットでは，図 2 に示すように車載カメラ画像中の歩行者の注視対象がアノテーションされている．加えて，対象となる歩行者の ID，BBox 座標，注視点座標，PEGO の BBox 座標，PEGO の物体クラス，歩行者の状態などのアノテーションが含まれている．

3. 歩行者の注視対象物推定

本節では，歩行者の注視対象物（PEGO）推定を行なう Transformer ベースの手法 “PEGO Transformer” を提案する．PEGO Transformer は図 3 に示す 4 つのモジュールで構成される．まず入力画像を Feature Extractor に入力して画像特徴量を獲得し，Deformable Transformer Decoder で物体ごとに対応する特徴を獲得する．獲得した物体ごとの特徴を，教師ラベルを生成する Label Generator と，物体間の注視対応関係を表現する Projection Layer にそれぞれ入力し，PEGO の推定を行なう．以下で各モジュールについて紹介する．

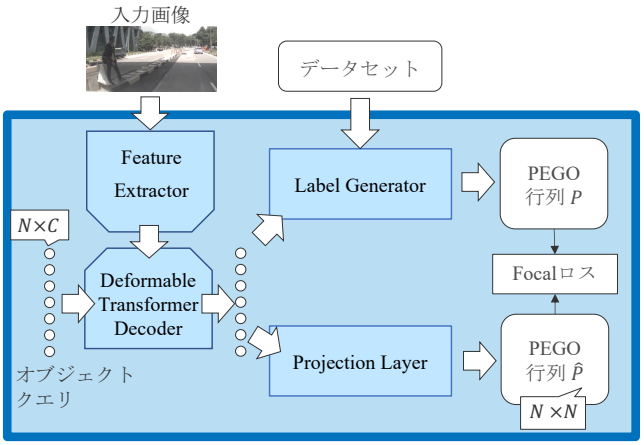


図 3: PEGO Transformer の構成

3.1 PEGO Transformer の構成

Feature Extractor. Feature Extractor は CNN バックボーンと Deformable Transformer Encoder [11] で構成され，画像特徴量の獲得を行なう．まず，入力画像 $\mathbf{x} \in \mathbb{R}^{C \times H \times W}$ を CNN バックボーン（ResNet [4]）に入力して画像特徴量を得る．次に，得られた画像特徴量を平坦化し，Positional Embedding を加えて Deformable Transformer Encoder に入力する．Deformable Transformer Encoder の self-attention モジュールを通して各画像特徴同士がそれぞれ影響し合う相互関係を捉える．

Deformable Transformer Decoder. Deformable Transformer Decoder [11] は，Feature Extractor の出力を入力とし，以下の手順を通して各物体に対応する特徴量とオブジェクトクエリ（$o_q \in \mathbb{R}^C$） [2] を対応付ける．まず，N 個の o_q，$\mathcal{O}_\mathcal{Q} = \{o_{q1}, o_{q2}, ..., o_{qN}\}$ を予めランダムな値で初期化し，Deformable Transformer Decoder に入力する．入力された $\mathcal{O}_\mathcal{Q}$ は，Deformable Transformer Decoder の cross-attention モジュールによって Feature Extractor の出力と物体に対応する特徴が対応付けられ，さらに self-attention モジュールを通してそれらの関係性を捉える．これにより，$\mathcal{O}_\mathcal{Q}$ の各 o_q はそれぞれ物体と 1 対 1 に対応した特徴量となる．

Projection Layer. Projection Layer は Transformer Encoder [9] と MLP で構成される．Transformer Encoder は，$\mathcal{O}_\mathcal{Q}$ を入力として物体間の注視対応関係を捉える．MLP は，Transformer Encoder の出力を後述する PEGO 行列 $\widehat{\boldsymbol{P}}$ に投影する役割を担う．

Label Generator. Label Generator は Projection Layer

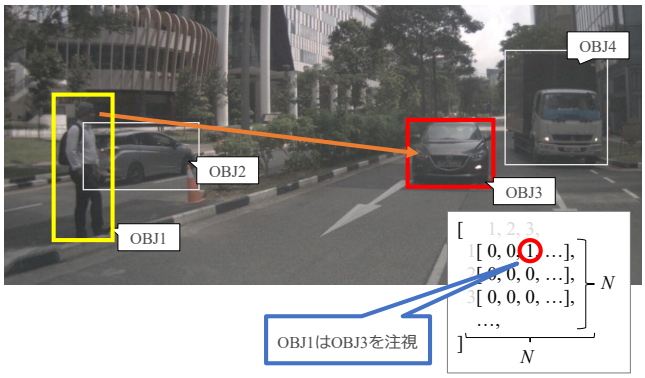


図 4: PEGO 行列の例

の訓練に用いる，教師ラベルの PEGO 行列 $\boldsymbol{P}$ を生成する．PEGO 行列 $\boldsymbol{P} \in [0,1]^{N\times N}$ は，画像中の物体間の注視関係を表現した $N \times N$ の行列である．i 番目の $\mathcal{O}_\mathcal{Q}$，$o_q i$ に対応する物体が，j 番目の $\mathcal{O}_\mathcal{Q}$，$o_q j$ に対応する物体を注視している場合，PEGO 行列の要素を $\boldsymbol{P}_{i,j} = 1$ とする．図 4 に PEGO 行列の例を示す．1 番目の $\mathcal{O}_\mathcal{Q}$（o_{q1}）に対応する歩行者が 3 番目の $\mathcal{O}_\mathcal{Q}$（$o_{q2} s$）に対応する車を注視しているとき $\boldsymbol{P}_{1,3} = 1$ となり，それ以外は 0 となる．また，歩行者以外に対応する $\mathcal{O}_\mathcal{Q}$ の行についてはすべて 0 となる．

教師ラベルの PEGO 行列 $\boldsymbol{P}$ を生成するために，まず N 個の o_q，$\mathcal{O}_\mathcal{Q}$ 中からそれぞれ物体領域を囲う矩形 (BBox) とクラスを推定する．そのうち，推定されたクラスが歩行者であり，かつその尤度がしきい値 δ_I 以上となる o_q を探す．これらは歩行者の検出結果として扱い，その o_q が $\mathcal{O}_\mathcal{Q}$ の何番目の要素であるか (インデックス) を $\mathcal{M} = \{m_1, m_2, \ldots\}$ として記憶する．次に，データセット中から各 $m \in \mathcal{M}$ に対応する歩行者の注視対象物のアノテーションを取得し，その物体が $\mathcal{O}_\mathcal{Q}$ の n 番目の要素に対応するとき，$\boldsymbol{P}_{m,n} = 1$ とする．

3.2 損失関数

PEGO 行列の値の多くは 0 となるため，単純な損失関数（例えば BCE 損失 [6]）はクラス不均衡の問題からモデルの学習ができない．そこで，クラス不均衡がある場合でも効率よく学習可能な，次式の Focal 損失 [5] を用いる．

$$\mathcal{L} = -\sum_{i}^{N}\sum_{j}^{N} \alpha(1-\widehat{\boldsymbol{P}}_{i,j})^{\gamma} \log(\widehat{\boldsymbol{P}}_{i,j}), \tag{1}$$

なお，α と γ は異なるサンプル数を持つクラスの損失のバランスを調整するハイパーパラメータである．

3.3 PEGO 推定

PEGO Transformer で求めた PEGO 行列 $\widehat{\boldsymbol{P}}$ から，PEGO の推定を行なう．ここで，$\widehat{\boldsymbol{P}}$ 中で歩行者に対応する行のインデックスを $\widehat{\mathcal{M}} = \{\widehat{m}_1, \widehat{m}_2, \ldots\}$ とする．そして，各 $m \in \widehat{\mathcal{M}}$ に対して次式により PEGO 候補に対応する $\mathcal{O}_\mathcal{Q}$ のインデックス θ_m を得る．

$$\theta_m = \arg\max_i \widehat{\boldsymbol{P}}_{m,i}. \tag{2}$$

そして，推定されたクラス尤度がしきい値 δ_{II} 以上である $\widehat{\theta}_m$ を得る．

$$\widehat{\theta}_m = \begin{cases} \theta_m & \text{if } f_{CP}(\mathcal{O}_\mathcal{Q}[\theta_m]) > \delta_{II} \\ \varnothing & \text{otherwise} \end{cases} \tag{3}$$

ここで $f_{CP}(\mathcal{O}_\mathcal{Q}[\theta])$ は，$\mathcal{O}_\mathcal{Q}$ の θ 番目の o_q に対応する物体のクラス尤度を表す．$\widehat{\theta}_m = \varnothing$ の場合，θ_m を除く PEGO 候補（$m \in \widehat{\mathcal{M}} \setminus \theta_m$）に対し，式 (2) と (3) を繰り返し適用する．これにより，対象歩行者 $m \in \widehat{\mathcal{M}}$ の PEGO に対応する添字 $\widehat{\theta}_m$ を得る．

4. 実験

訓練済みの PEGO Transformer を用いて，PEGO の推定を行なった．本節ではまず実験設定を示し，次に実験結果について述べる．

4.1 実験設定

我々が構築した歩行者注視対象物データセット [12] を用い，歩行者 652 人分のデータに対して交差検証を行なった．データ拡張として，入力画像の左右反転と，Projection Layer に入力するオブジェクトクエリの順番のシャッフルを行なった．

学習時には，Projection Layer のパラメータのみを更新対象とした．Feature Extractor と Deformable Transformer Decoder は，Deformable DETR [11] で訓練済みの重みを用いて初期化した．使用するオブジェクトクエリの数は $N = 40$ とした．δ_I と δ_{II} はともに 0.3 に設定した．Focal 損失のハイパーパラメータ α は 0.25，γ は 2 に設定した。

4.2 結果

表 1 に PEGO 推定の精度を示す．歩行者それぞれに対応する PEGO の推定結果で最も尤度が高いものが正解と一致するかを Top1 正解率として評価に用いた．また，上位 2 位，3 位，4 位，5 位で推定した PEGO 中で，我々のデータセットと一致するかどうかを評価した結果を Top2, Top3, Top4, Top5 正解率とし，あわせて示している．

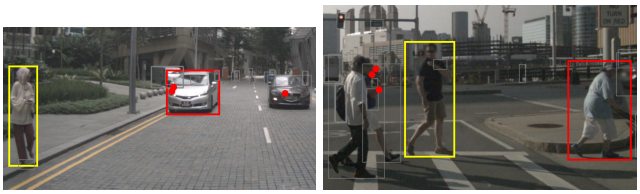
図 5(a) は PEGO 推定の成功例である．これらの結果から，従来の注視推定手法が必要としていた頭部特徴に依存せず，PEGO の推定に成功したことが確認できる．一方，図 5(b) は PEGO 推定に失敗した例である．対象歩行者の男性は後ろを振り向いているが，それとは反対側の女性を PEGO と誤推定した．

5. 考察

PEGO Transformer の有効性を検証するため，損失関数

表 1: 歩行者注視対象物（PEGO）推定の正解率

	ロス関数	Top1(%)	Top2(%)	Top3(%)	Top4(%)	Top5(%)
チャンスレート		3.44	6.78	10.0	13.1	16.2
BCE ロスを使用	**BCE**	20.4	30.2	42.6	46.7	49.4
式 (3) による絞り込みなし	Focal	19.8	23.8	25.4	28.9	38.4
PEGO Transformer	**Focal**	**24.7**	**32.3**	**45.6**	**51.6**	**62.8**



(a) 正解: 女性は中央の車を注視.

(b) 不正解: 男性は後方の男性を注視しているが右側の女性を注視と誤推定.

図 5: PEGO の推定例：対象歩行者は黄 BBox，推定した PEGO は赤 BBox で示す．赤点はデータセットに収録された注視点.

の変更としきい δ_{II} を変更した際の性能の変化を評価した．

損失関数を Focal 損失から BCE 損失に変更した場合，PEGO 推定精度が低下することを確認した．PEGO 行列では，歩行者の注視対象物体に対応する位置のみ値が 1 となり，それ以外はすべて 0 となる．そのため，値 0 と値 1 の割合が大きく異なり，クラス間に不均衡が生じる．その結果，BCE 損失は非常に小さい値となり，学習が進まなかったと考えられる．このことから，本モデルにおいてはクラスの不均衡を考慮した Focal 損失が重要であったといえる．

次に PEGO 推定において，しきい値 $\delta_{II} = 0$ として，式 (3) による絞り込みを行なわないと，精度が低下することを確認した．δ_{II} を小さくすると物体ではない誤検出も PEGO 推定の対象に含まれることが原因だと考えられる．今後，Projection Layer の学習時にこのような誤りを考慮することで，精度が向上することが期待される．

6. むすび

本発表では，歩行者の注視対象物（PEGO）を推定する Transformer ベースの手法を提案した．PEGO Transformer は，従来の注視対象推定手法で用いられる頭部位置の情報を明示的に用いることなく PEGO 推定を可能とした．今後の課題として，骨格情報を考慮した PEGO Transformer の提案，などが挙げられる．

謝辞本研究の一部は JSPS 科研費 23H03474 による．本研究の一部は名古屋大学のスーパーコンピュータ「不老」の一般利用にて実施した．

参考文献

[1] Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G. and Beijbom, O.: nuScenes: A multimodal dataset for autonomous driving, *In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11618–11628 (2020).

[2] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. and Zagoruyko, S.: End-to-End object detection with transformers, *In Proceedings of the European conference on computer vision*, Springer, pp. 213–229 (2020).

[3] Hata, R., Deguchi, D., Hirayama, T., Kawanishi, Y. and Murase, H.: Detection of distant eye-contact using spatio-temporal pedestrian skeletons, *In Proceedings of the IEEE 25th International Conference on Intelligent Transportation Systems*, pp. 2730–2737 (2022).

[4] He, K., Zhang, X., Ren, S. and Sun, J.: Deep Residual Learning for Image Recognition, *In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016).

[5] Lin, T.-Y., Goyal, P., Girshick, R., He, K. and Dollár, P.: Focal Loss for Dense Object Detection, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 42, No. 2, pp. 318–327 (2020).

[6] Liu, L. and Qi, H.: Learning Effective Binary Descriptors via Cross Entropy, *In Proceedings of the 2017 IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1251–1258 (2017).

[7] Recasens, A., Khosla, A., Vondrick, C. and Torralba, A.: Where are they looking?, *In Proceedings of the Advances in Neural Information Processing Systems*, Vol. 28, pp. 199–207 (2015).

[8] Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z. and Anguelov, D.: Scalability in Perception for Autonomous Driving: Waymo Open Dataset, *In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2443–2451 (2020).

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u. and Polosukhin, I.: Attention is All you Need, *In Proceedings of the 2017 Advances in Neural Information Processing Systems*, Vol. 30 (2017).

[10] Wang, B., Hu, T., Li, B., Chen, X. and Zhang, Z.: GaTector: A Unified Framework for Gaze Object Prediction, *In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19588–19597 (2022).

[11] Zhu, X., Su, W., Lu, L., Li, B., Wang, X. and Dai, J.: Deformable DETR: Deformable Transformers for End-to-End Object Detection, *In Proceedings of the 9th*

International Conference on Learning Representations (2021).

[12] 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 コンピュータビジョンとイメージメディア研究会，Vol. 2023-CVIM-233, No. 57, pp. 454–459 (2023).