大規模画像言語モデルを活用した Coarse-to-Fine 型セマンティックシーン検索

立花 卓遠[1,2,a)]　坂口 翼[1,b)]　大西 正輝[1,c)]　櫻田 健[1,3,d)]

概要

本研究では、空間動画像データベース (SAI-DB) に対し、テキストによる高精度かつ高速なシーン検索を実現するため、Coarse-to-Fine 型セマンティック検索システムを提案・実装した。具体的には、Coarse ステップでは、ユーザのシーンを記述したクエリ文と、画像から生成された説明文をそれぞれベクトルに変換し、類似度を計算して候補を絞る。そして、Fine ステップにおいて、GPT-4o を用いた VQA を利用することで、インタラクティブにクエリと一致する画像を取得できる。評価実験として、新宿駅付近で撮影されたシーン動画像に基づいてタスクを設定し、提案システムの有効性を定性的に確認した。

1. はじめに

屋内外のシーン動画像データを収集する技術や機器の発達、そしてオープンデータの増加に伴い、自動運転、XR[*1]、都市計画やモニタリングの分野でシーン動画像データが活用され始めている [1]。例えば、膨大なシーン動画像を深層学習を用いて処理することにより、土地分類や人流計測、市街状況の把握などが可能になる [11]。現状では、物体検出や意味的領域分割を通じてシーン画像のタグ付けやクラスタリングを試みる研究が多い [8]。しかしながら、実際の屋内外のシーン動画像にはさらに豊かな情報が含まれており、タグ付けやクラスタリングの結果として得られる限られた情報とは差異がある。現在の技術で「物体が存在するかどうか」や「どの領域にそれが存在するか」といった単純化された情報は抽出できるものの、「物体がどのような状態に置かれているか」や「街の活気や雰囲気」といった、より高次のシーン理解は達成されていない。

OpenAI の GPT シリーズに代表される大規模画像言語モデルの急速な発展に伴い、モデルを追加学習させることなく、シーン動画像からより多くの情報を抽出することが可能になりつつある。さらに、プロンプト次第で、より人間の理解に近い形でシーン理解を記述し、特定のタスクに適したシーン画像の説明文を得ることもできる。(例: 街中の点字ブロックに関する問題点を探るために、「写真中に写っている点字ブロックの状況を詳述して」というプロンプトで大規模言語モデルに説明文を生成させる。)

図 1 空間動画像データベース (SAI-DB) の UI。図中の「段階的検索」が Coarse-to-Fine 型セマンティック検索に該当する。

このような、大規模画像言語モデルのシーン理解によって得られた説明文をシーン動画像と共に地理座標系に登録した空間動画像データベースが SAI-DB である（図 1）[4]。図 1 に示されているように、SAI-DB にはシーン画像をテキストクエリする機能があり、クエリ文に合った地点をインタラクティブに検索することができる。坂口ら (2024) [4] が実装した手法では、クエリ文が、登録されている説明文の一部と「完全一致」した地点のみを選択し表示するようにしていた。このシンプルな手法でも検索自体は可能であるが、同じ情景を思い浮かべていても、クエリ文は個々人でばらつきがあり、期待するクエリに辿り着くまでに多くの試行を必要とする場合もある。

一方で、クエリ文のばらつきに対応した、より堅牢な意味的検索（セマンティック検索）を実現するために、テキストの埋め込みベクトルを作成し、それらの類似度によってデータを取得するシステムが様々な分野で研究されて

1 産業技術総合研究所
2 東京大学
3 京都大学
a) t-tachibana@g.ecc.u-tokyo.ac.jp
b) sakaguchi-tsubasa@aist.go.jp
c) onishi-masaki@aist.go.jp
d) sakurada@i.kyoto-u.ac.jp
*1 Cross Reality, VR・AR 技術の総称

いる [2, 10]。これに倣い、SAI-DB においても、クエリ文や説明文を全て埋め込んでベクトル類似度でクエリする「曖昧検索」機能を実装したが、説明文が長くなると、埋め込みベクトルも多義的になり、正確なクエリが困難であった。この課題を解決するために、本論文では、いくつかの検索システムで用いられている Coarse-to-Fine のアプローチ [3, 9] を採用し、SAI-DB に Coarse-to-Fine 型セマンティック検索（「段階的検索」）機能を統合した。

2. 関連研究

杉本 (2020, 2022) [6, 7] らが提案しているムービーマップは、動画像データを地理座標系に登録し、ユーザがマップ上で街中を快適に散策し、没入感のある体験を楽しめるようにすることを目的としている。そのためには各交差点での移動方向をスムーズに接続させる必要があり、近年の研究では特に、交差点の識別方法に工夫が凝らされている。2020 年の研究では Visual SLAM 技術を活用した交差点識別を実装し [6]、また、2022 年の研究では、PDoT 法による交差点の自動識別手法が導入され、より効率的かつ精度の高い交差点識別が可能となった [7]。

一方で、本論文で取り扱う SAI-DB は、動画像を地理座標系に登録したデータベースという点ではムービーマップと同じ分類に属するが、システムの目的が異なる。本論文では、街中のシーン画像をインタラクティブにテキストクエリできることに焦点を当てており、ユーザが特定のシーンや場所を迅速かつ直感的に検索できる機能を重視している。このように、ムービーマップが快適な街中散策の提供を主眼としているのに対し、SAI-DB はタスクに応じたシーン画像の効率的な検索と利用を目的としている点で差別化されている。

また、Coarse-to-Fine のアプローチを採用したテキストクエリの研究として、Text2Pos と Text2Loc が挙げられる。Kolmet ら（2022）[3] は、テキストクエリから 3D 点群データ内の位置を特定する Text2Pos のタスクに取り組んだ。この研究では、Coarse のステップで、テキストとサブマップの埋め込みベクトル類似度で top-k のサブマップを取得し、Fine のステップにおいて、テキストクエリと候補サブマップ内のオブジェクトをマッチングさせている。Xia ら（2023）[9] による Text2Loc ではこれをさらに発展させ、Fine のステップで、テキスト情報と点群データを異なるブランチで処理した後に、それらを融合して位置を推定することでクエリ精度を改善した。

Text2Loc と Text2Pos の、膨大な地理的データに対するテキストクエリにおいて、精度と計算量のバランスを取るために Coarse-to-Fine を採用している点では、本論文で提案する Coarse-to-Fine 型セマンティック検索と一致している。一方で、Text2Pos や Text2Loc が正確な地点の特定を目的としているのに対し、本論文のシステムはタスクに応じたシーン画像の取得を目的としている。また、Coarse ステップにおいて、ベクトル類似度を使って top-k の候補を取得する点では共通しているが、本論文の手法では、Fine ステップにおいて、取得した top-k のシーン画像とクエリ文を GPT-4o に渡して適切なものを選択させるため、そもそも存在しない不適切なクエリを排除することができるという点で差別化されている。

3. 提案手法

本論文ではベクトル類似度による top-k 検索と大規模画像言語モデルの VQA を組み合わせた Coarse-to-Fine セマンティックシーン検索を提案する（図 2）。

3.1 SAI-DB に登録されているデータ

SAI-DB には以下のデータが地理情報と共に登録されている。

- 360° カメラで撮影した equirectangular 形式のシーン画像
- 大規模画像言語モデル (GPT-4o) で作成したシーン説明文
- 大規模言語モデル (text-embedding-3-large) で作成したシーン説明文埋め込みベクトル

シーン説明文を作成するときは、equirectangular 形式のシーン画像を cubemap 形式に変換して GPT-4o に渡している。例えば、図 2 で示されている説明文としては、以下のプロンプトで生成した json ファイルのうち “雰囲気” の項目を登録されている。なお、実際のプロンプトでは one-shot 方式で例を与えることにより、出力される json ファイルの形式を固定している。

> あなたの仕事は、360 度のパノラマビューから撮影された 4 つの画像について説明することです。添付画像は、正面、左、背面、右の視点を示すように切り取られています。添付画像のシーンについて、下記の項目を json 形式で出力してください。
> ・シーンの雰囲気を説明してください。
> ・何をするのに適した雰囲気の場所か説明してください。
> ・適する項目の度合いを 10 段階でスコア化してください。

また、シーン説明文から、OpenAI の text-embedding-3-large を使用してシーン説明文埋め込みベクトルを作成し、ベクトルデータベースに登録した。ベクトルデータベースとしては Spotify 社が開発した Voyager [5] を使用した。

3.2 Coarse ステップ

ユーザがクエリ文を入力すると、OpenAI の text-embedding-3-large を使用してクエリ文を埋め込み、シー

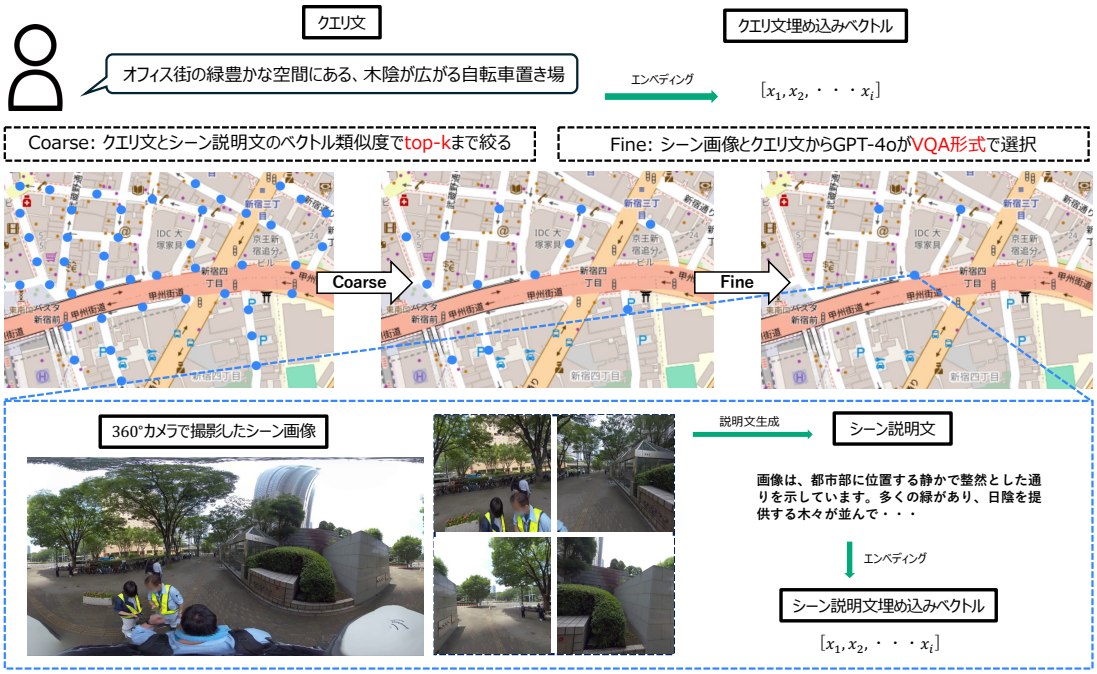

図 2　Coarse-to-fine 型セマンティックシーン検索の概要

ン説明文埋め込みベクトルとのコサイン類似度で top-k のシーン画像を取得する。

3.3　Fine ステップ

クエリ文と Coarse ステップで取得してきたシーン画像を GPT-4o に渡し、クエリ文と合致する画像を VQA 形式で選択させる。具体的には以下のプロンプトで VQA を行っている。

> { クエリ文を代入 } というクエリに一致する画像を選んで、クエリと合致する順に並び替えてください。また、その写真がどうしてクエリと一致すると考えたのか、詳述してください。クエリに適合しないと考える画像は説明不要なので回答から除外してください。

3.4　提案手法の優位性

提案手法では、膨大なシーン画像データに対するクエリを一定の精度を保ちながら迅速に実行するため、Coarse-to-Fine アプローチを採用している。特に SAI-DB はインタラクティブなシステムであるため、応答速度はユーザエクスペリエンスの評価に直結する。シーン画像の説明文が詳細であることから、埋め込まれるベクトルは多義にわたり、coarse ステップのベクトル類似度だけではクエリの精度が不十分である。Fine ステップにおいて、大規模画像言語モデルによる VQA を併せて行うことにより、高精度なクエリを実現した。さらに、Fine ステップで明らかに不適切なシーン画像を除外することができるので、存在しない状況に対するクエリに対しても堅牢性を持つ。また、Fine ステップにおいて、シーン画像を選択した理由について大規模画像言語モデルに説明させることで、選択理由の透明化を図ることもできる。

4.　実験

4.1　データセット

SAI-DB に登録されているシーン動画像は、バックパック型の MMS(Leica Pegasus) に 360° カメラ (Insta360 Pro2) を搭載し、新宿駅周辺の地上および地下街でそれぞれ約 20km を徒歩または自転車で撮影して取得した。この動画から、20m 間隔で 1595 枚のシーン画像データを取得し、それぞれに対して GPT-4o で説明文生成を行った。

4.2　定性的結果

提案手法で示したプロンプトおよび処理で登録した説明文に対する、Coarse-to-Fine セマンティックシーン検索の実行例を図 3 に示す。図 3 の例では、「オフィス街の緑豊かな空間にある、木陰が広がる自転車置き場」というクエリ文を用いて、Coarse ステップで top-k($k = 10$) として実行した。結果として、Fine ステップで 3 つの画像に絞り込まれた。

この例は以下の点で、セマンティック検索の成功を示している。

- 高次的な特徴を捉えていること
- 街の雰囲気や特定の特徴を持った駐輪場を理解できていること
- Fine 検索によって適切に絞り込まれていること

図 3 Coarse-to-Fine 型セマンティックシーン検索の例。「オフィス街の緑豊かな空間にある、木陰が広がる自転車置き場」のクエリ結果。

例えば、Fine で絞られた画像で、選択理由として「特に木陰が自転車置き場に影響を与えている様子が見受けられる」と記述されてものがあり、これは高次的な特徴を捉えた理想的なクエリである。

一方で、駐輪場に木陰がなく、道端の木陰と混同してしまっている例も見受けられる。また、Fine ステップで選ばれなかった画像にも、木陰のある自転車置き場が存在してしまっており、今後の改善が求められる。

5. おわりに

本論文では、SAI-DB に対し、Coarse-to-Fine 型セマンティックシーン検索を提案・実装した。これにより、ユーザはテキストを用いて高精度かつ高速にシーン画像を検索することが可能となった。評価実験として、新宿駅付近で撮影されたシーン動画像に基づいたタスクでの例を示した。

今後の課題としては、定量評価を行い、提案システムの性能をさらに詳しく評価する必要がある。また、登録データ数の増加に伴うスケーラビリティの検証や、ユーザエクスペリエンスの改善も進めていく予定である。

6. 謝辞

本研究は、JST さきがけ（JPMJPR22C4）の支援を受けたものである。

参考文献

[1] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R. and Schiele, B.: The cityscapes dataset for semantic urban scene understanding, *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223 (2016).

[2] Covington, P., Adams, J. and Sargin, E.: Deep neural networks for youtube recommendations, *Proceedings of*

the 10th ACM Conference on Recommender Systems, ACM, pp. 191–198 (2016).

[3] Kolmet, M., Zhou, Q., Ošep, A. and Leal-Taixé, L.: Text2pos: Text-to-point-cloud cross-modal localization, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6687–6696 (2022).

[4] Sakaguchi, T., Onishi, M. and Sakurada, K.: テキスト検索可能なシーン画像データベースの構築, *Technical Report on Computer Vision and Image Media (CVIM)*, Vol. 2024, No. 26, pp. 1–6 (2024).

[5] Spotify: Voyager: A Vector Database, `https://spotify.github.io/voyager/`. Accessed: 2024-06-22.

[6] Sugimoto, N., Ebine, Y. and Aizawa, K.: Building Movie Map-A Tool for Exploring Areas in a City-and its Evaluations, *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 3330–3338 (2020).

[7] Sugimoto, N., Ikehata, S. and Aizawa, K.: Intersection Prediction from Single 360° Image via Deep Detection of Possible Direction of Travel, *arXiv preprint arXiv:2204.04634* (2022).

[8] Wang, F., Yuan, C., Li, J. and Liu, B.: What makes a place special? Research on the locality of cities in the Yellow River and Rhine River Basins based on Street View Images, *Indoor and Built Environment*, Vol. 31, No. 2, pp. 435–451 (2022).

[9] Xia, Y., Shi, L., Ding, Z., Henriques, J. F. and Cremers, D.: Text2loc: 3d point cloud localization from natural language, *arXiv preprint arXiv:2311.15977* (2023).

[10] Zhang, H., Wang, S., Zhang, K., Tang, Z., Jiang, Y., Xiao, Y. and Yang, W. Y.: Towards personalized and semantic retrieval: An end-to-end solution for e-commerce search via embedding learning, *In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2407–2416 (2020).

[11] Zhang, Y., Zhang, F. and Chen, N.: Migratable urban street scene sensing method based on vision language pre-trained model, *International Journal of Applied Earth Observation and Geoinformation*, Vol. 113, p. 102989 (2022).