ISS-A-048

画像と質問文によるインタラクティブマルチモーダル屋内現在位置推定

李 欣耘† 古田 諒佑†† 入江 豪††† 山本 洋太† 谷口 行信†
† 東京理科大学 †† 東京大学 ††† NTT コミュニケーション科学基礎研究所

1. はじめに

近年，駅などの屋内施設の複雑化により， GPS に代わる屋内位置推定技術の需要が増加している．この需要に対応するために，画像認識を用いた現在位置推定手法が提案された．この手法は，事前に撮影した位置情報付きの画像（参照画像）をデータベースに保存しておき，ユーザが撮影した画像（クエリ画像）をデータベース内の参照画像と照合し，類似画像を検索することで，現在位置を推定する．本研究では，類似画像が多く存在している屋内でも，高精度かつ少ない手間で現在位置を探索するため，質問文を用いる屋内現在位置推定手法を提案する．具体的には，参照画像から質問を自動生成し，ユーザとのやり取りから，現在位置を絞り込んでいく手法である．

2. 従来手法

従来の現在位置推定手法として，局所特徴量[1]と，CNN 大域特徴量に基づく手法[2]が提案されているが，類似物体が多い屋内での精度が低いという問題がある．そこで，我々はクエリ画像を前後左右の4方向に増やした多視点画像と，それに対応する距離計算手法[3]を提案した．しかし，クエリ画像4枚の撮影に手間がかかり，周囲の人のプライバシへの配慮が必要となる問題がある．

本研究では，この問題を解決し，精度を維持したまま撮影回数を削減するために，ユーザに景観に関する質問応答を要求するアプローチを検討する．

3. 提案手法

提案手法では，通常の画像検索に，周辺に見える物体の有無を問う質問応答を組み合わせる（図1）．まず，(1)ユーザが撮影した一枚のクエリ画像に類似する参照画像を検索し，位置候補を絞る．(2)すべての参照画像から，物体検出器を用いて，物体(机，椅子等)を検出し，「(物体ラベル)に近いですか？」という質問を自動に生成する．(3)効率よく位置候補を絞り込むために質問を選択する．(4)ユーザと質問応答を行いながら，現在位置を特定する．これによって，撮影枚数が減少し，なおかつ，ユーザからインタラクティブに情報を取得することができる．

ステップ(3)において，少ない質問回数で，正解を導き出すために，最小の条件付き entropy を持つ質問を求める:

$$H_{Q_i} = -\sum_{j\in\{\text{Yes, No}\}}\sum_{k} P(A_k)\log_2\frac{P(A_k)}{\sum_l P(A_l)}. \qquad (3.1)$$

ここで，$P(A_k) = 1\left[ans_i^k = j\right]\left(e^{S_k}/\sum_{j=1}^{K} e^{S_j}\right)$である．また，$1\left[ans_i^k = j\right]$は参照画像に対して質問$Q_i$の答えが$j \in \{\text{Yes, No}\}$の時1を，それ以外は0を取る指数関数である．$S_k$は参照画像を類似度順に並べた時の$k$番目(Top k)の類似度，$A_k$は Top k が正解となる事象である．最小となるQ_i

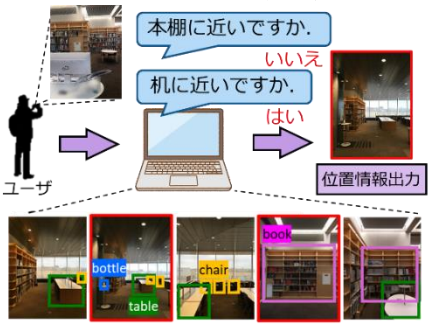

図 1: 画像と質問文を用いた現在位置推定の流れ

表 1: 提案手法と従来手法の比較

手法		方向数	Accuracy [%]
多視点画像距離[3]		1	73.08
		4	84.02
提案手法	full-auto	1	63.78
	semi-auto	1	74.68

をユーザに尋ね，ユーザの返答から，結果を絞り込み，やり取りを繰り返す．

4. 実験

実験方法 公開データセット West Coast Plaza(WCP) Dataset[4]を用いる．1m 間隔で撮影した参照画像1264枚と，ランダムの場所で撮影したクエリ画像312枚がある．物体検出器は COCO で学習済みの YOLOv3[5]を用いた．36種類のラベルのうち，屋内施設に存在する物体ラベル（机，椅子等）34種を残し，それ以外のラベル（人，犬）2種を削除した．

結果・考察 誤差1m以内で正解したクエリ画像の割合 One-Meter-Level Accuracy を評価指標とする．表1に示す通り，物体検出誤りの影響で提案手法(full-auto)の精度は従来手法[3](方向数1)よりも低いが，物体検出のクラスラベルを修正した結果(semi-auto)，精度は1.60ポイント向上した．よって，物体検出の精度が向上すれば，質問応答で撮影の手間を減らしながら高精度な屋内位置推定を実現することが期待できる．

5. 今後の課題

今後は，物体検出器のファインチューニングや情景内文字認識との組み合わせを検討し，精度改善を図る．

参考文献

[1] James Philbin, et al. Object Retrieval with Large Vocabularies and Fast Spatial Matching. In *CVPR*, pp. 1–8, 2007.
[2] Filip Radenović, et al. Fine-Tuning CNN Image Retrieval with No Human Annotation. *IEEE TPAMI*, Vol. 41, No. 7, pp. 1655–1668, 2019.
[3] Xinyun Li, et al. Accurate Indoor Localization Using Multi-View Image Distance. In *IEVC*, 3A-2, 2021.
[4] Meng-Jiun Chiou, et al. Zero-Shot Multi-View Indoor Localization via Graph Location Networks. In *ACM Multimedia*, pp. 3431-3440, 2020.
[5] Farhadi, Ali, and Joseph Redmon. Yolov3: An Incremental Improvement. In *CVPR*, 2018.