2ZE-04

不完全な物体検出結果に基づく対話を通じた目的地点推定のための質問選択

時末 卓幹 * 武田 龍 * 駒谷 和範 * 翠 輝久 † 細見 直希 ‡ 山田 健太郎 ‡
* 大阪大学 産業科学研究所 †ホンダ・リサーチ・インスティチュート・USA ‡株式会社本田技術研究所

1. はじめに

近年では，システムがユーザの指示発話を物理世界の状況に基づいて理解する技術の必要性が高まっている．例えば自動運転タクシーにおいて，ユーザの周辺の環境情報に基づいた指示を理解することで，より柔軟な目的地点の設定が実現できることが期待される．

本研究では物理世界の状況の認識結果として，図1のような物体検出結果を利用する．これは，領域，種別，色の3つの属性を持つ検出物体の集合である．

物理世界ではユーザの指示する可能性のある，あらゆる物体を検出できるわけではないため，物体検出結果は不完全である．例えば図1で,「カラオケ屋の看板」をユーザは認識できるが，この物体は物体検出結果には含まれないことが多い．ユーザの指示するあらゆる物体を検出するのは，計算やデータ作成のコストが高く，困難である．そこでシステムとユーザの認識できる物体の違いに，対話を通して対処する．

本研究では目的地点を推定するための対話における，質問選択手法を提案する．システムはユーザに質問を行い，応答から目的地点の情報を得る．ただし単純な質問を続けるだけでは目的地点の情報がほとんど得られず，効率よく推定できない．そのため，質問を適切に選ぶ必要がある．なお対話を通してユーザの指示する物体を推定する研究は存在するが [1]，本研究は物体検出結果が不完全な場合に着目している．

2. 問題設定

本稿では街中の画像において，テキスト対話を通して，ユーザの指示する目的地点を推定するシステムを考える．目的地点の推定とは，画像に対する物体検出結果から，目的地点に最も近い検出物体を推定することとする．システムは大きく分けて「質問選択」「発話解釈」の2つのモジュールから成る．本研究ではこの「質問選択」モジュールの設計を提案する．

「質問選択」モジュール (図2上段) では効率よく推定を進めるために，ユーザに対する質問を適切に選択する．ここで効率がよいとは，より少ない発話数で，目的地点により近い検出物体を推定できることとする．本稿における質問選択は「○○は近くに見えますか」などの質問テンプレートにおいて「○○」に入れる検出物体を選択することとする．検出物体は a_i と表記する．

「質問選択」モジュールの入力は，t 回目のやり取りの時，候補 $O_t = \{a_1, a_2, \ldots, a_n\}$ とする．n は候補の要素数であり，$|O_t| = n$ とする．この候補をより絞り込める応答を得られるように質問を選択する．検出物体 a_i はその領域の中心座標 (x_i, y_i)，種別，色を属性として持つ．a_i の種別がバッグや服など，人の所有物である場合，a_i の座標は所有者の領域の中心座標とする．

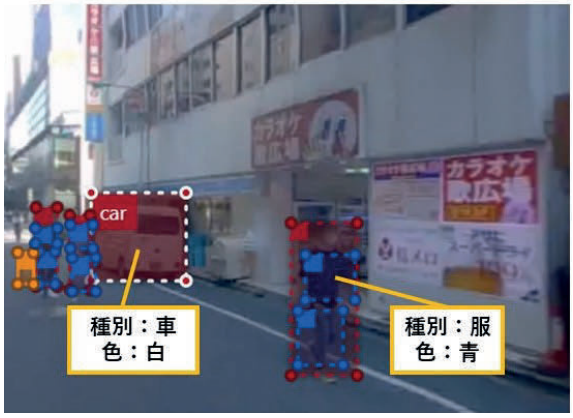

図 1: 物体検出結果の例

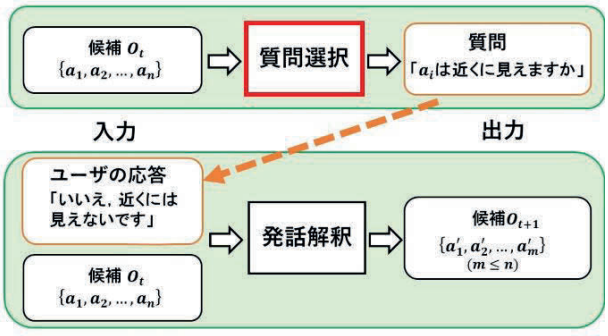

図 2: システムの構成と入出力 (t 回目のやり取り)

選択した質問に対するユーザからの応答を得て,「発話解釈」モジュール (図2下段) は現在の候補 O_t を絞り込む．絞り込みの結果，候補 O_{t+1} が出力される．

$|O_{t+1}| = 1$ となるまで,「質問選択」「発話解釈」の順に処理を繰り返す．最後に候補に残った検出物体の座標が，目的地点の推定座標となる．なお対話はユーザ発話から始まるので，対話開始時は「発話解釈」モジュールのみ使用し，初期候補 O_0 とユーザ発話から，O_1 を得る．$|O_1| = 1$ でない場合，$O_1 \leftarrow O_0$ とする．

3. 2分探索的な質問選択

効率的に目的地点を推定するため，検出物体の空間的な位置関係を用いた2分探索的な手法を提案する．本手法では初期候補 O_0 を，検出物体の内，種別・色の2つで他の検出物体と区別がつくもののみとする．

2分探索的手法に基づいて質問を選択する．t 回目のやり取りでの候補を $O_t = \{a_1, a_2, \ldots, a_n\}$(ただし $x_1 < x_2 < \ldots < x_n$) とする．候補を2つの集合に分割するため，まず $a_i \in O_t$ の x 座標の平均 $\bar{x}_t = (\sum_{i=1}^{|O_t|} x_i)/|O_t|$ を求める．2つの集合 L_t, R_t を，$x_p \leq \bar{x}_t < x_{p+1}$ となる p により $L_t = \{a_1, a_2, \ldots, a_p\}, R_t = \{a_{p+1}, a_{p+2}, \ldots, a_n\}$ とする．ここで L_t と R_t のどちらに目的地点に最も近い

Question selection through dialogues for goal estimation based on incomplete object detection: *Takumi Tokisue, Ryu Takeda, Kazunori Komatani (Osaka Univ.), Teruhisa Misu (Honda Research Institute USA, Inc.), Naoki Hosomi, and Kentaro Yamada (Honda R&D Co., Ltd.)*

Copyright ©2022 Information Processing Society of Japan.
All Rights Reserved.

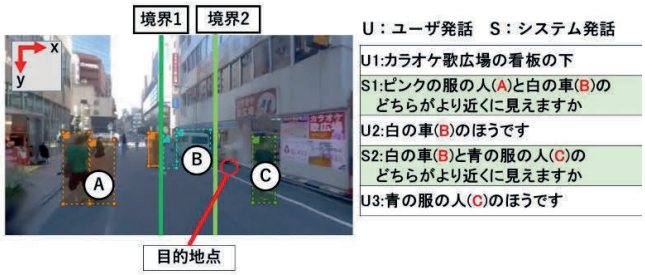

図 3: 提案手法に基づく目的地点推定対話例

a_i が存在するか分かれば，候補を一方の集合に絞ることができる．そのために $a_l \in L_t, a_r \in R_t$ を用いて，「a_l と a_r のどちらがより近くに見えますか」という質問を行う．a_l がより近いと言われた場合，$O_{t+1} \leftarrow L_t$ とする．一方で a_r がより近いと言われた場合，$O_{t+1} \leftarrow R_t$ とする．今回は $a_l \leftarrow a_p, a_r \leftarrow a_{p+1}$ とした．

a_l, a_r がごく近い場合，近傍処理と呼ぶ処理を行う．2つの検出物体 a_i, a_j の座標間の距離を D_{ij} とする．この距離がある閾値 ϵ 以下，つまり $D_{ij} \leq \epsilon$ の場合に a_i, a_j がごく近いとする．$D_{lr} \leq \epsilon$ の場合，$|x_{p-1} - \bar{x_t}| < |x_{p+2} - \bar{x_t}|$ なら $a_l \leftarrow a_{p-1}$，$|x_{p-1} - \bar{x_t}| > |x_{p+2} - \bar{x_t}|$ なら $a_r \leftarrow a_{p+2}$ とする．また t 回目のやり取りで $D_{ij} \leq \epsilon (\forall a_i, \forall a_j \in O_t)$ の場合，目的地点の推定座標は候補 O_t の平均座標とする．

図 3 に対話例を示す．画像中の赤い丸の中心が，推定すべき目的地点である．画像中で領域に囲まれた 6 つの物体が，初期候補 O_0 となる．今回は最初のユーザ発話 (U1) で候補は絞れず，$O_1 \leftarrow O_0$ となった．システムはまず $x = \bar{x_1}$ となる境界 1 を求める．次に a_l, a_r を求める．図 3 では $a_l \leftarrow (A), a_r \leftarrow (B)$ となった．この時に出力された質問が S1 である．ユーザの応答 (U2) から，$(B) \in R_1$ が目的地点により近いと分かり，$O_2 \leftarrow R_1$ とする．同様に $x = \bar{x_2}$ となる境界 2 を求め，$a_l \leftarrow (B), a_r \leftarrow (C)$ として質問を行う (S2)．ユーザの応答 (U3) から $O_3 \leftarrow R_2$ とすると，$O_3 = \{(C)\}, |O_3| = 1$ となり，(C) が最も目的地点に近いと推定できる．

4. 実験

4.1 実験条件

本実験では画像に対し，ユーザの応答を模したユーザモデルとシステムとで対話シミュレーションを行い，提案手法の性能を評価した．

使用する画像は計 107 枚で，解像度は全て 1080 × 1920 ピクセル (以降 px) である．各画像にはシステムが推定すべき目的地点の座標を事前に 1 つずつ人手で設定した．

ユーザの応答は以下のモデルを仮定してシミュレーションを行った．モデルは設定された目的地点をもとに，システムの質問に対しては適切に応答するものとした．具体的には，目的地点との距離が ϵ 以内の検出物体は目的地点の近くとして，500 px 以上の検出物体は見えないとして応答した．またモデルは 1 発話目に，画像ごとに用意した 5 つの発話から，ランダムに 1 つを与える．この発話は予備実験において，各画像で設定した目的地点を指すために，5 人のユーザが実際に発話したものである．

システムがユーザ発話から認識可能な物体は，種別が MSCOCO[2] および Image Net[3] から抜粋した 122 種に 3 種を加えた計 125 種類であるものとした．また物体の色は 13 種類が認識可能とした．本実験では各画像の，これらに該当する物体に人手でアノテーションしたものを，画像に対する物体検出結果として用いた．

本実験では提案手法である「2 分探索+近傍処理」に加え，次の 3 つの手法を比較に用いた．どの手法も初期候補 O_0 の設定は提案手法と同じで，提案手法を含むすべての手法で $\epsilon = 200$ px とした．

- Random
  候補 O_t からランダムに 1 つずつ a_i を選び，目的地点に近いか質問する．質問した a_i が近くにあると応答された場合，$O_{t+1} = \{a_i\}$ とする．近くにないと応答された場合，$O_{t+1} = O_t - \{a_i\}$ とする．
- Random + 近傍処理
  Random において，質問した a_i が目的地点の近くになかった場合，a_i とともに $D_{ij} \leq \epsilon/2$ となる $a_j \in O_t$ も候補から除く．
- 2 分探索
  提案手法で，近傍処理を行わない．

評価指標は，推定までに交わした平均発話数，最大発話数，および推定座標と目的地点との平均距離とした．対話内容はユーザの 1 発話目などに依存するため，各手法で 107 枚の画像それぞれにおいて対話を行い，これを 100 回繰り返した．評価指標は，これらの対話で得られるものの平均である．なお推定が完了しなかった場合は評価指標の算出には含めなかった．107 枚の画像の内，推定が完了した画像数の平均を推定完了平均画像数とした．

4.2 実験結果と考察

表 1 に実験結果を示す．平均発話数，最大発話数，目的地点との平均距離はそれぞれ値が小さいほど性能が良い．

平均発話数は「2 分探索+近傍処理」が最も良い結果で，Random に比べ約 18%少ない．また最大発話数は Random と比べ半分以下に抑えられた．

一方で目的地点との平均距離は，「2 分探索+近傍処理」が最も大きかった．これは候補の平均座標を，目的地点の推定座標として出力する場合があるためだと考えられる．ただし他の手法との差は最大でも 25px，画像の短辺との比率で 2.3%程度と，軽微である．

表 1: 実験結果

手法	平均発話数	最大発話数	目的地点との平均距離 [px]	推定完了平均画像数
Random	4.33	27	148	92
2 分探索	4.56	11	157	105
Random +近傍処理	4.08	23	147	92
2 分探索 +近傍処理	3.57	11	173	102

5. おわりに

今回の実験から画像中の目的地点推定において，提案手法は最大発話数を抑えやすく，効率的であることが分かった．本稿では物体検出結果の位置のみを利用して質問を選択したが，物体検出結果の集合内での属性の分布の利用も今後検討する．実際に人と対話した際の検証も今後の課題のひとつである．

参考文献

[1] Felix Gervits, et al. Decision-theoretic question generation for situated reference resolution: An empirical study and computational model. In *Proc. ICMI*, pp. 150–158, 2021.

[2] Tsung-Yi Lin, et al. Microsoft COCO: Common Objects in Context. In *Proc. ECCV*, pp. 740–755, 2014.

[3] Jia Deng, et al. Image Net: A large-scale hierarchical image database. In *Proc. IEEE CVPR*, pp. 248–255, 2009.

Copyright ©2022 Information Processing Society of Japan.
All Rights Reserved.