画像特徴とテキスト特徴を用いた画像ツイートの位置推定

松尾　真[1]　柳井　啓司[1]

1. はじめに

近年スマートフォンやタブレット端末の普及により、Twitter の利用者が急増している。ジオタグ付画像ツイートは、Twitter の視覚情報と位置情報を扱う研究において、重要なサンプルデータとなる。しかし、大量に存在する画像ツイートの中で、実際に位置情報が付与されたものは非常に少ない。したがって、大量のジオタグなし画像ツイートの位置推定を行うことで、研究におけるこれらの利用価値を大いに高めることができると期待される。

2. 目的

本研究では、ジオタグのない画像ツイートの位置を画像特徴とテキスト特徴を用いて推定することを目的とする。画像特徴からの位置推定には DCNN(Deep Convolutional Neural Network) を用いた類似画像検索を、テキストからの位置推定には GeoNLP [*1]を用いた地名語抽出を使用する。

3. 関連研究

渡辺ら [1] は「Twitter を用いた実世界ローカルイベント検出」の研究において、テキストを用いたツイートの位置推定を行った。

Hays らは IM2GPS [2] の研究において大量の位置情報付き画像を利用した類似画像検索によって単一画像の位置推定が可能であることを証明した。

Babenko ら [3] は DCNN 特徴が画像検索アプリケーションにおいても有用であることを示し、PCA などの手法で低次元化させた場合もその精度が低下しないことを実証した。

4. 提案手法

4.1 提案手法の流れ

本研究は画像ツイートに対して以下の手順で位置推定を行う。

(1) 画像による位置推定
(2) テキストよる位置推定
(3) 推定された位置情報の統合

4.2 従来の手法

従来の手法では画像による位置推定には特徴点マッチングによる類似画像検索を、Nister ら [4] が考案した Vocaburaly Tree で高速化して使用した。今回の手法では画像表現として DCNN 特徴を用いることで類似画像検索の精度の向上を図る。

4.3 画像による位置推定

位置情報付き学習画像データセットに対する DCNN 特徴による類似画像検索を用いて、M 枚の類似画像を選出し、それらの位置情報と類似度から推定位置にスコアを与える。

本研究では Overfeat [*2]を用いて抽出した 4096 次元の DCNN 特徴を用いる。また、類似画像検索を高速・低容量化するため、本研究では PCA を用いて DCNN 特徴を 64 次元まで圧縮した。

4.4 テキストよる位置推定

テキストによる位置推定には、入力画像ツイートのテキストから、GeoNLP を用いて地名と位置情報を取り出すことで位置推定を行う。また、今回の実験では日本語ツイートのみを扱うものとする。

4.5 推定された位置情報の統合

画像、テキストそれぞれの情報から複数の推定位置候補とスコアを抽出し、図 1 のように地図を緯度経度によって離散化したグリッドにスコアを与え、最もスコアの高いグリッドを求める。テキスト特徴と画像特徴、2 つの情報から得られたグリッド L_i のスコアをそれぞれ $P_t(L_i|I)$, $P_v(L_i|I)$ で定義し、式 1 で統合する。w_t, w_v はそれぞれのテキスト特徴と画像特徴情報に対して与える重みである。

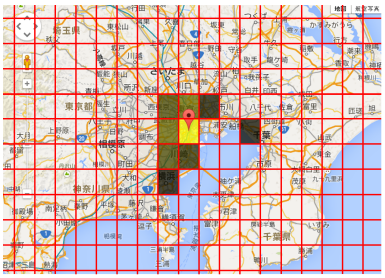

図 1　地図の離散化

$$P(L_i|I) = \frac{w_v P_v(L_i|I) + w_t P_t(L_i|I)}{\sum_{k=1}^{n} w_v P_v(L_k|I) + w_t P_t(L_k|I)} \quad (1)$$

最終的に最もスコア $P(L_i|I)$ が高いグリッドの中心を最終的な推定位置とする。

4.6 信頼度に基づく画像の重み付け

また、実験では w_t と w_v の値は固定で 5 パターン扱う他、w_t と w_v を自動決定するパターンも扱う。テスト画像 I による位置推定を信頼できるかどうかの尺度を信頼度とし、$B(I)$ で表す。$B(I)$ は類似画像が位置推定の材料となるならば同一のグリッドに集中するという仮説に基づいて求める。テスト画像 I の類似画像の上位 N 枚の所属するグリッドを調べ、その内同じグリッドに属するものの枚数の最大値を K とし、信頼度は以下の式 2 で求め、w_v と w_t の値は式 3 で決定する。

$$B(I) = \frac{e^{\frac{K}{N}} - 1}{e - 1} \quad (2)$$

$$w_v = B(I), w_t = 1 - B(I) \quad (3)$$

[1] 電気通信大学

[*1] http://agora.ex.nii.ac.jp/GeoNLP/

[*2] http://cilvr.nyu.edu/doku.php?id=software:overfeat:start

5. 実験

5.1 実験データと評価方法

実験には表 1 ような 2011 年から 2014 年の位置情報付画像ツイートを実験データとし、w_t と w_v を A F の 6 パターンに分け、類似画像数 M=50, パターン E において使用する類似画像数 N=10 で行った。

表 1 実験データ

	学習データ	テストデータ
データの数	約 240 万	4000
テキストの利用	無	有
位置情報の利用	有	無

評価はテストデータの内、正解位置との地球球面上の距離が一定距離以内であるものの割合で行う。地球球面上の距離は球面三角法を用いて求める。

5.2 実験結果

表 2 は M=50 の際の、従来の手法 (BoF) と今回の手法 (DCNN) での結果である。

表 2 実験結果（M = 50, 単位は%）

	feature	w_t	w_v	5km	10km	50km	100km
A	BoF	1.00	0.00	36.0 (1440 枚)	57.2 (2288 枚)	65.9 (2636 枚)	68.3 (2732 枚)
B		0.75	0.25	35.8 (1432 枚)	57.5 (2300 枚)	67.3 (2692 枚)	69.7 (2788 枚)
C		0.50	0.50	35.3 (1412 枚)	56.8 (2272 枚)	66.6 (2664 枚)	68.8 (2752 枚)
D		0.25	0.75	31.8 (1272 枚)	50.6 (2024 枚)	58.6 (2344 枚)	60.5 (2420 枚)
E		0.00	1.00	2.6 (104 枚)	6.0 (240 枚)	13.7 (548 枚)	16.0 (640 枚)
A	DCNN	1.00	0.00	36.0 (1440 枚)	57.2 (2288 枚)	65.9 (2636 枚)	68.3 (2732 枚)
B		0.75	0.25	36.7 (1468 枚)	58.7 (2348 枚)	67.2 (2688 枚)	69.9 (2796 枚)
C		0.50	0.50	36.6 (1464 枚)	58.3 (2332 枚)	66.6 (2664 枚)	69.4 (2776 枚)
D		0.25	0.75	35.0 (1400 枚)	55.2 (2208 枚)	62.8 (2512 枚)	65.4 (2616 枚)
E		0.00	1.00	4.1 (164 枚)	8.1 (324 枚)	15.9 (636 枚)	18.0 (720 枚)
F		AUTO	AUTO	36.3 (1452 枚)	59.1 (2364 枚)	68.9 (2756 枚)	71.4 (2856 枚)

固定のパターン（A〜E）については、テキスト特徴のみ (パターン A) に比べ、画像特徴のみ (パターン E) の精度は非常に低いという傾向は変化しなかったが、従来の手法よりも、B〜E の精度にわずかな上昇が見られた。特に精度が高かったのはパターン B, パターン C であり、どちらもすべての範囲においてパターン A の精度を上回る結果となった。また、従来手法と比べて最も精度の向上が大きかったのはパターン D であり、最大 5 %程度の上昇が見られた。

自動重み付けを行ったパターン F の精度は全ての範囲で従来手法の結果を上回っていた。また、同手法の固定パターンと比較しても、ほとんどの範囲で上回る精度を出しており、全体を通して最も優れた精度であると言える。

6. 考察

実験結果から、画像特徴のみではテキスト特徴のみの場合に比べ、精度が低かったが、画像特徴とテキスト特徴を統合した際の精度はテキスト特徴単体の時よりも大きくなることが確認できた。また、類似画像検索の手法を特徴点マッチングから DCNN 特徴に変えたことによる、画像特徴を用いる位置推定の精度向上も確認できた。

また、図 2、図 3 はそれぞれ画像特徴のみで位置推定を行った際（パターン E）の成功例（誤差 5km 以内）と失敗例（誤差 300km 以上）の画像である。図 2 は京都の墨染駅と東京タワーの風景であり、図 3 は仙台駅内の風景と奈良駅の屋外の風景である。前者が一目で分かるランドマークであるのに対し、後者は室内の特定物体だけが映っていたり、周囲の風景が判別しにくいものであったりと、視覚情報から位置情報への繋がり薄いことが分かる。これらの例から、テストデータの内、画像特徴単体で成功しやすいものは不特定多数の人間が同じ風景を撮影するランドマーク画像であり、失敗しやすいものは位置情報との関連性が薄いものであると言える。

これらの画像のパターン F における位置推定を行ったところ、信頼度は前者は共に 0.8 を上回り、逆に後者は共に 0.1 を下回っており、推定位置は前者は変わらず、後者は誤差 10km 近くまで改善された。したがって、今回の自動重み付けの手法は画像の信頼度の計測法として有用であったと言える。

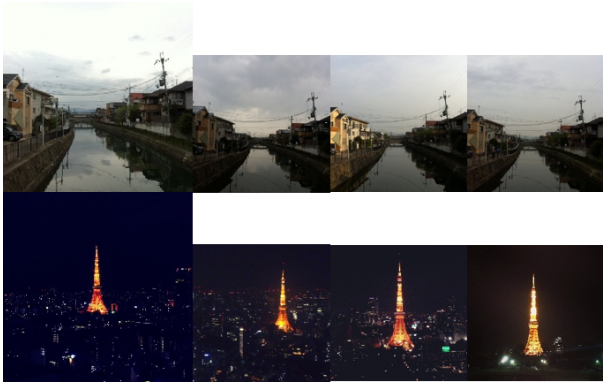

図 2 成功例 (左端：テスト画像, その他:学習画像)
京都墨染駅 (推定位置:京都府京都市伏見区, 誤差:1.76km)(上)
東京タワー (推定位置:東京都港区, 誤差:1.70km)(下)

図 3 失敗例 (左端：テスト画像, その他:学習画像)
仙台駅 (推定位置：東京都港区, 誤差:315km)(上)
奈良駅 (推定位置：東京都港区, 誤差:373km)(下)

参考文献

[1] K. Watanabe, M. Ochi, M. Okabe, and R. Onai. Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs. In *Proc. of the 20th ACM international conference on Information and knowledge management*, pp. 2541–2544, 2011.

[2] J. Hays and A. A. Efros. IM2GPS: estimating geographic information from a single image. In *Proc.of IEEE Computer Vision and Pattern Recognition*, 2008.

[3] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky. Neural codes for image retrieval. In *Proc.of European Conference on Computer Vision*, 2014.

[4] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Proc.of IEEE Computer Vision and Pattern Recognition*, 2006.