Twitter Photo Geo-Localization Using Both Textual and Visual Features

Shin Matsuo, Keiji Yanai

国立大学法人 電気通信大学

Introduction

- Twitter and Weibo
  - timelines and on-the-spot-ness
  - include much information on various events
- Geotagged photo tweets
  - Has locations where photo were taken
  - Geotagged photo tweets is very limited

Introduction

- Objective
  - localizing a Twitter photo using both textual features and visual features
- localization from texts
  - GeoNLP 1
- localization from visual features
  - image search for a geotagged photo database
  - SIFT or DCNN features

Related Work

- Watanabe et al.
  - Estimate locations of tweets from texts
- Hays et al (IM2GPS)
  - image retrieval for a large-scale geotagged image

Proposed Method
A. Overview

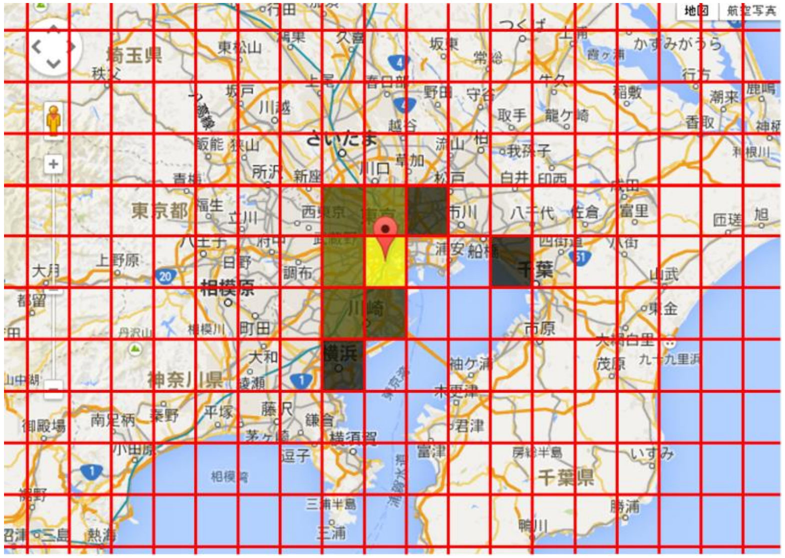
1) Location estimation by visual features

2) Location estimation by Twitter messages texts.

3) Integration of the locations estimated by the two kinds of features

B. Grid-based location estimation

- grid-based location rather than a pair of longitude and latitude
- We evaluate possible grids by giving scores, and select the grid with the best score as the final estimated location.



C. Twitter photo localization by visual features

- Photo locations with image retrieval for a large-scale geotagged image database
  - several millions of geotagged photo tweets
- Features
  - SIFT feature
  - DCNN feature (Overfeat)
    - 4096d -> 64d by PCA

C. Twitter photo localization by visual features

- Image retrieval for a database
  - Top M similar images for a given image
  - The visual-feature-based score

$$S_v(L_i|I) = \sum_{j=1}^{M} \frac{1}{\sqrt{j}} \phi(E_j - i) \qquad P_v(L_i|I) = \frac{S_v(L_i|I)}{\sum_i S_v(L_i|I)}$$

    - E_j represents the location grid index of j-th retrieved images
    - $\phi(x) = 1(x = 0), 0(x \neq 0)$

D. Text-based location estimation

- GeoNLP
  - Extracts place names such as Tokyo and New York
  - Estimate location based on the dictionary
- The textual-feature-based score of i-th grid

$$S_t(L_i|I) = \sum_{j=1}^{N} \phi(E_j - i) \qquad P_t(L_i|I) = \frac{S_t(L_i|I)}{\sum_i S_t(L_i|I)}$$

  - E_j represents the location grid index of j-th retrieved images
  - $\phi(x) = 1(x = 0), 0(x \neq 0)$

E. Integration of estimated location

- Textual score $P_v(L_i|I)$
- Visual score $P_t(L_i|I)$

- Integrated score

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

F. Automatic weight estimation

- reliable score $B(I)$
  - represents how extent the estimated locations to image I concentrate to one grid

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

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

  - K represents the number of the estimations in the grid

Experiments

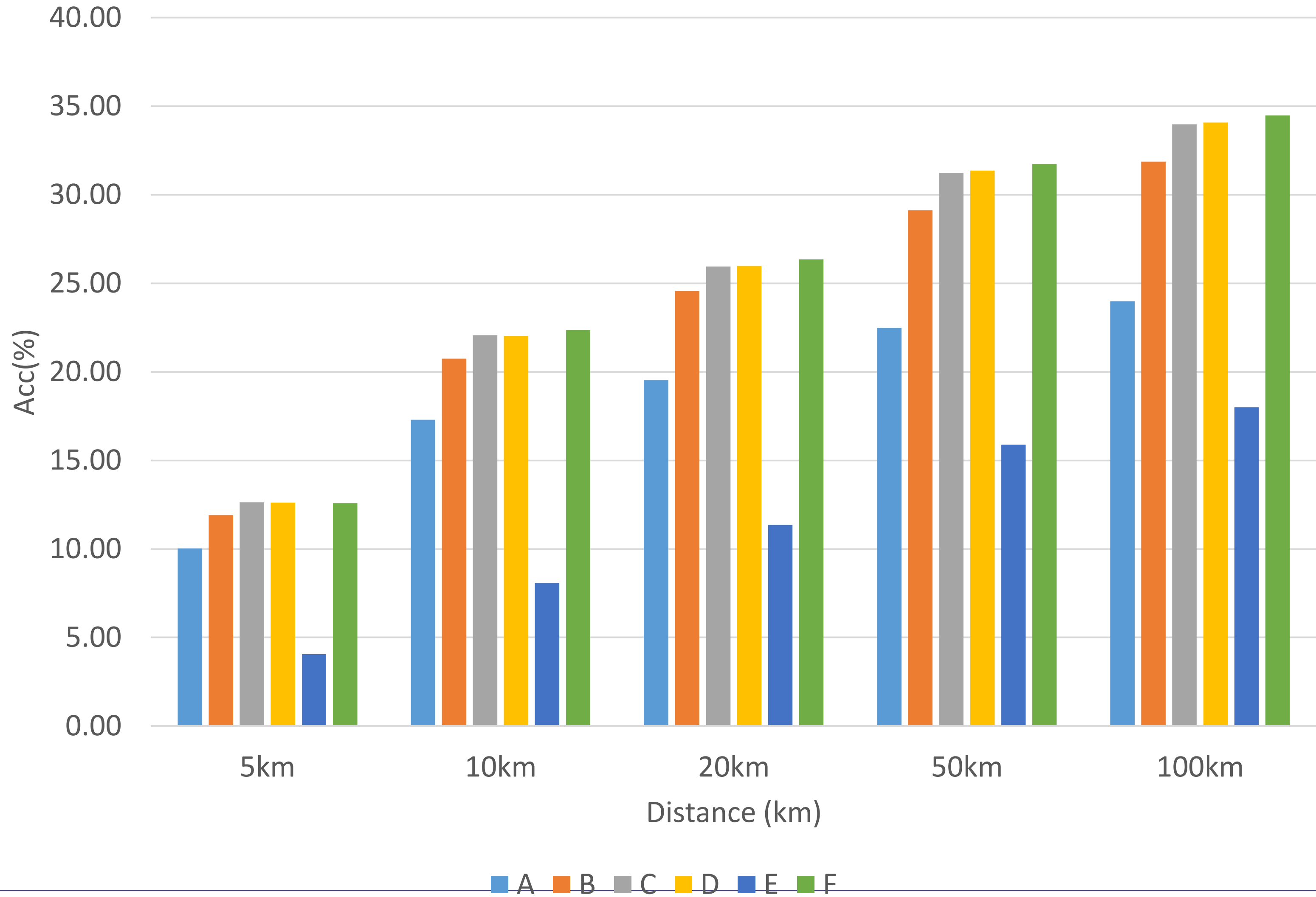
- Dataset
  - Training data
    - 2014/01~2015/01
    - About 240,000
  - Test data
    - 2011/02~2014/12
    - Around 4,000
  - Similar image number:M=50
  - Grid size : 0.1°　(about 10km)

TABLE II

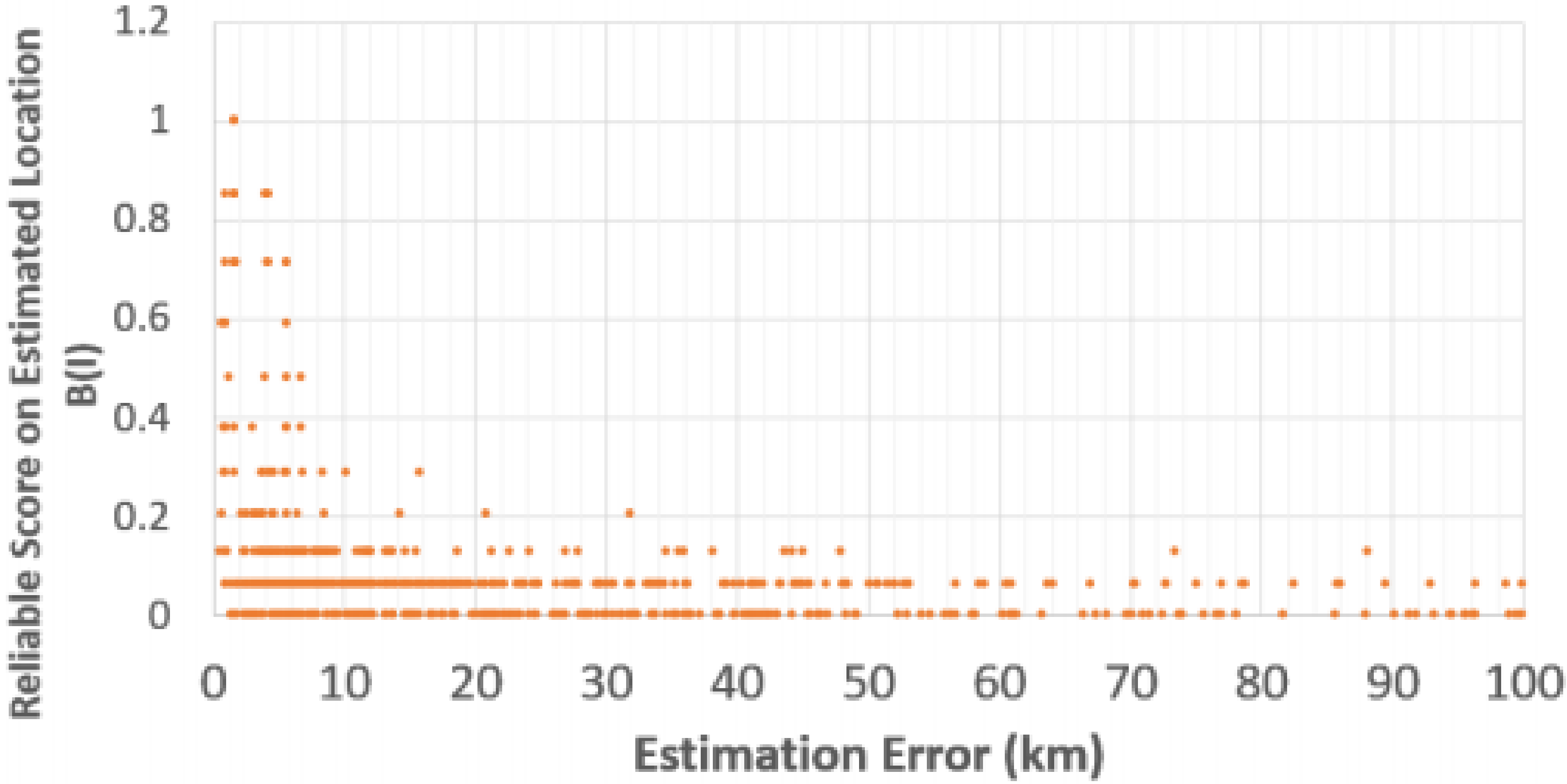
LOCALIZATION ACCURACY (%) WITH $M = 50$.

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

Experiments



Experimental results



Examples

Examples

- Visual features only

Conclusion

- We proposed a method to localize Twitter photos
- integration of both features improved localization accuracy compared to using only single modality