LARGE-SCALE TWITTER FOOD PHOTO MINING AND ITS APPLICATIONS

Keiji Yanai *Kaimu Okamoto* *Tetsuya Nagano* *Daichi Horita*

Department of Informatics, The University of Electro-Communications, Tokyo
1-5-1 Chofugaoka, Chofu-shi, Tokyo, 182-8585, JAPAN

ABSTRACT

Many people are posting photos as well as short messages to Twitter every minutes from everywhere on the earth. By monitoring the Twitter stream, we can obtain various kinds of photos with texts which help understand the current state of the world visually. Since 2011, we have been continuously collecting photos from the Twitter stream for about eight years. We are collecting food photos as well as generic geotagged photos, since we are intensively working on multimedia processing on foods. . In this paper, we focus mainly on Twitter food photos. Because foods are one of the most popular contents of Twitter photos, we can collect a large number of food images from Twitter. In fact, we have collected more than two million food photos so far. In this paper, we present the analysis on the food photos collected from the Twitter stream. In addition, we describe some applications using Twitter photos including world food photo analysis and food photo translation/generation.

Index Terms— food photo mining, Twitter photo mining, Twitter stream, food image recognition, food image generation, GAN

1. INTRODUCTION

Nowadays, posting photos to SNSs such as Facebook, Instagram and Twitter is a part of everyday activities in human life. Such photos are not only photos taken at special situations such as trip photos and party photos but also usual photos taken in everyday life like foods and documents. A great large number of people are posting their photos with short messages to SNSs every minutes from everywhere on the earth. People usually send their photos to SNSs right after taking the photos on the spot. Therefore, by monitoring the photo stream, we can get to know the current state of the world visually. This is very different from static image databases and the results of Web image search engines. Twitter is a promising data source of SNS photos, since we can gather Tweet photos from the stream in an almost real-time way using the Twitter streaming API, whereas Instagram does not provide the way to collect photos from the real-time stream, and most of the photos on Facebook can be seen among only "friends".

In [1], we proposed a real-time geotagged tweet photo mapping system, "WorldSeer", which visualizes photo tweets with geotags on the Google Maps in a real-time way as well as stores information on geo-photo tweets to our database continuously by monitoring the Twitter streaming via the Twitter Streaming API. We have been collecting both geo-photo tweets and photo tweets without geotags continuously since February 2011. In 2019, the system is still keeping on collecting photos from Twitter. Figure 1 shows some geotagged photos of "ramen" on "WorldSeer". Unfortunately, due to the change of the service policy of the Twitter stream for unpaid users, the number of tweets per unit time were dramatically reduced to one tenth in May 2015. Before then, we collected five to ten million geotagged images per month, and after then we collected five to ten hundred thousand geotagged images per month. However, totally we have collected three hundred and twenty one million geotagged photos so far. The large-scale geotagged photo database we have created is expected to be very useful resource for multimedia research.

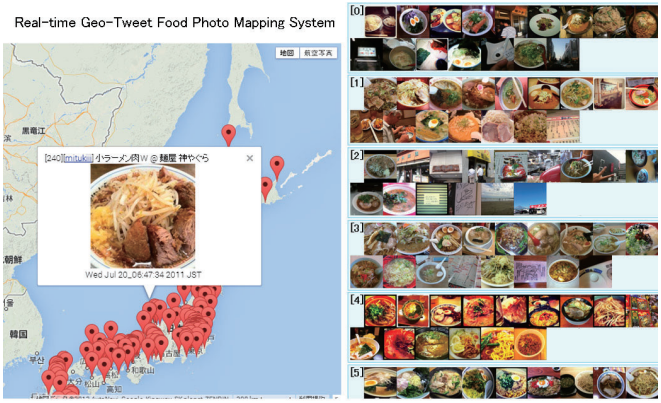

Fig. 1. Food images on the real-time geotagged Twitter photo mapping system, "WorldSeer". The locations of the collected geotagged food photos are marked on the Google map

.

In fact, using the Twitter photo database we have been creating for eight years, we have done many works including food photo mining [2], event photo mining [3, 4, 5], geo-location estimation of non-geotagged photos [6], Twitter world photo analysis [7, 8], food image translation [9, 10], real-time food image translation for VR [11, 12], and food image generation [13, 14].

In this paper, we introduce our works on Twitter photos especially related to food domain. At first, we explain the current food photo mining system which has been updated

from [2] and report the number of food photos we have collected so far. Next, we introduce Twitter world photo analysis [7] and world food photo analysis [8]. As applications of Twitter food photo mining, we briefly explain food photo translation [9, 10], and generation [14].

2. TWITTER FOOD PHOTO MINING

In this section, we explain the Twitter food photo mining system, which started running in 2011 and is still running in our lab until now. The current system has been updated compared to the version of the food photo mining system reported in [2]. At that time, the system employed a conventional image recognition method which consists of HOG, Fisher Vector, and SVM, while the current system employs a CNN-based food classier [15].

By combining keyword-based search and food image recognition, we mine food photos from the Twitter stream. To collect food photos from Twitter, we monitor the Twitter stream to find the tweets containing both food-related keywords and photos. If found, we download the image instantly and apply a 101-class food classifier which can classify a given image into one of the pre-defined 100 food categories or non-food category with 0.03 seconds. The 100 food categories are the same as the food categories in the UEC-FOOD100 dataset [16][1] which consists of 100 kinds of foods commonly eaten in Japan.

As a food image classifier, in the current system, AlexNet [17] is still being used, since the performance by combining keyword search achieved almost perfect at the time of 2015 as mentioned later. We pre-trained AlexNet with 2000 ImageNet categories including ILSVRC 1000 categories and additional 1000 food-related categories selected from the 21841-category full ImageNet. In general, it is effective to pre-train a CNN with the dataset augmented with additional categories related to the target recognition task. After pre-training, we fine-tune the pre-trained AlexNet. For Twitter food mining, it is required to exclude non-food photos. To do that, we add a non-food category to the 100 categories of UEC-FOOD100. We used 10000 non-food photos collected from Twitter as training data for a non-food category. We fine-tune the pre-trained AlexNet as a 101-class classifier which can recognize non-food photos as well as 100-class food photos.

Table 1. Food classification performance on UEC-FOOD100.

classifier	top-1 rate	top-5 rate
FV (Color+HOG)	65.32	86.70
AlexNet(1000)	75.25	93.19
AlexNet(2000)	78.48	94.85
AlexNet(1000+ex)	76.68	94.40
AlexNet(2000+ex)	**78.77**	**95.15**
DenseNet [18]	83.9	97.1
WideResNet [19]	89.6	99.2

[1] http://foodcam.mobi/dataset/

Table 1 shows the evaluation results for UEC-FOOO100. FV(Color+HOG) corresponds to the classifier employed in the original system. AlexNet(1000) and AlexNet(2000) represent the performance of AlexNet pre-trained with 1000 ILSVRC ImageNet and 2000 ImageNet containing 1000 food-related categories, respectively. By adding 1000 food-related categories, the top-1 accuracy was improved by 3%. AlexNet(1000+ex) and AlexNet(2000+ex) represent the results by the AlexNet fine-tuned with the augmented UEC-FOOD100 we created by adding at most 1000 food photos mined from Twitter to each of the 100 categories. In the actual system, AlexNet(2000+ex) is being currently used. For reference, the performance on UEC-FOOD100 by DenseNet [18] and WideResNet [19] are shown in the table as ones by the state-of-the-art CNNs.

Following the previous report [2], for evaluation, we used 122,328,337 photo tweets with Japanese messages out of 988,884,946 photo tweets over all the world collected from May 2011 to August 2013 for two years and four months from the Twitter Stream. From these photo tweets, we selected 1,730,441 photo tweets the messages of which include any of the name words of the 100 target foods as the first step.

In the previous version, as the second step, they applied a "foodness" classifier (FC) to all the images selected by keywords. After applying FC, we applied 100-class one-vs-rest individual food classifiers. As a result, we obtained 470,335 photos which were judged as food photos corresponding to any of the 100 target food categories by the processing pipeline described in [2]. In the pervious version, we adopted Fisher Vector and linear classifiers for FC and 100-class classifiers.

Instead of FC and FV-based 100-class food classifiers, we applied the 101-class AlexNet-based CNN classifier, which can achieve non-food photo detection and food photo classification simultaneously, to 1,730,441 Twitter photos selected by keyword search of the food names. In this large-scale food classification experiment, we found that CNN was very suitable for large-scale image data, since it takes only 0.03 seconds to classify one food photo with GPU and totally it needed about four hours to classify 1,730,441 photos by four GPU machines. Finally, we obtained 581,271 food photos, which was 1.24 times as many as the result in the previous report.

Table 2 show the results of the top five categories and two additional categories out of 100 food categories , and show 40 automatically detected photos of each of "ramen noodle", "dipping noodle (tsukemen)", "sushi" and "omelet" in Figure 2. Note that the precision rates shown in the table were estimated by subjective evaluation of random sampled 1000 photos for each categories, and the rightmost column of Table 2 shows the number of the food photos detected by DCNN from the Twitter stream from May 2011 till now (July 17th, 2019) for about eight years. All the ranking and the number of collected photos of 100 food categories are shown in Table 3

Compared CNN with FC+100 which corresponds to the final results of [2], the number of obtained food photos and precision are improved. Especially the number of ramen pho-

Table 2. The number of selected photos and their precision(%) with four different combinations.

food category	raw	FC	FC+100	CNN	CNN(2019/7/16)
ramen noodle	275,652 (72.0%)	200,173 (92.7%)	80,021 (99.7%)	132,091 (99.5%)	500,210
beef ramen noodle	861 (94.3%)	811 (99.0%)	555 (99.7%)	590 (100%)	3,434
curry	224,685 (75.0%)	163,047 (95.0%)	59,264(99.3%)	68,091 (100%)	209,391
cutlet curry	10,443 (92.7%)	9,073 (98.0%)	6,339 (99.3%)	7,024 (99.9%)	23,401
sushi	86,509 (69.0%)	43,536 (86.0%)	25,898 (92.7%)	22,490 (99.8%)	130,501
dipping noodle	33,165 (88.7%)	24,896 (96.3%)	22,158 (99.0%)	22,004 (100%)	96,482
omelet with fried rice	34,125 (90.0%)	28,887 (96.3%)	17,520 (99.0%)	20,039 (99.9%)	12,859

tos were increased greatly, while the number of sushi photos were decreased. Although the precision of sushi in [2] was low, it was improved much and became almost perfect. This is because non-food photos representing inside sushi restaurants and people face photos were completely excluded by food-nonfood classification of the CNN. Regarding other foods than sushi, the precision rates were almost perfect. Only several photos are found in the 1000 random sampled photos in the evaluation time. We show some irrelevantly recognized photos in Figure 3.

3. WORLD PHOTO ANALYSIS

The tendency of photo contents on SNSs are different from region to region. This are expected to come from the difference of cultures, climates, people's interests and so on. Especially the content of the photos posted to SNSs is affected by people's thoughts on the privacy at SNSs which are expected to depend on their culture and history. In fact, the people in Eastern Asian such Japan and Korea do not like to post human face photos to SNSs, while Western and South-Eastern-Asian people posts many human face photos including selfy. To make this fact clear and explore it deeper, we analyzed geotagged Twitter photos by classifying them into five kinds of rough categories, "people", "building", "document", "scene" and "food". As results, we found there are quite large differences on regional tendency of posted Twitter photos regarding the photo genre distributions.

The existing work on SNS photo analysis such as [3, 4, 5] depends on texts attached to photos in general, because the number of SNS photos is so large like from millions to billions. The analysis using textual information attached to photos was effective for photo sharing sites such as Flickr. Since many of the Flickr users want to have many people see their photos, they tend to attach keywords or tags which expresses the content of photos directly to make it easier to search for their photos. On the other hand, attached texts to photos in SNSs such as Twitter and Instagram do not tend to represent the contents of photos directly in general. This is because the objective of attached messages are not for search but for explaining additional information which cannot be understood by just seeing photos. Therefore, in this work, we only image features without no textual information at all. This is possible by running CNN-based feature extractors on GPUs. Even one GPU can extract CNN-based features from million-scale images within one day.

3.1. Method

In this work, we analyze the regional tendency of rough categories of Twitter photos such as food, people and scenes using geotagged Twitter photos. As a target data, we use two million geotagged Twitter photos which we had gathered for half years in 2016. First, we extract 4096-d CNN features from all the images, and compress them into 128-d compressed features via PCA for reducing computational cost. Next, we cluster the images by K-means and classify only large clusters which have more than one hundred images into one of five typical rough categories or out of them. Finally, we compare the ratios of five categories between eight regions over the world.

3.2. Analysis on World Twitter Photo

For regional analysis, we used the nine regional divisions: East Asia, North America, South America, Europe, Africa, Middle East, South Asia, South-East Asia, and Oceania, as shown in Figure 4. Note that China was excluded from the target regions, since Twitter is prohibited to use in the Chinese region.

After clustering, we obtain clusters of the images which were semantically similar to each other. To analyze the tendency of posted photos, we classify the obtained photo clusters into one of the pre-selected photo genres. As photo genres, we use "people", "building", "document", "scene" and "food". These genres are decided based on the observation of clustering results of each of the nine regions. After genre classification of clusters, we compare the genre distributions of the images in each region regarding all the region to clarify the difference of regional tendency of Tweet photos.

By using five photo genres, we analyze the regional tendency of posted photos. We classify clusters by hand into one of the five genre. Although we can build a classifier to do that automatically, at this time we regard accuracy as most important rather than fully automatic processing. Thus, we check one by one to exclude noisy clusters which contain multiple genre images or no images corresponding to one of the five genres. In addition, sometimes there are clusters within which almost all the image are identical, which means a kind of spam image clusters. We also excluded them by hand as well.

Table 3. The ranking of Twitter photos of 100 foods.(2019/07/16)

1	ramen noodle	500,210
2	beef curry	209,391
3	sushi	130,501
4	omelet with fried rice	103,199
5	dipping noodles	96,482
6	pizza	89,568
7	jiaozi	67,196
8	okonomiyaki	61,919
9	beef steak	60,885
10	hambarg steak	53,588
11	toast	51,973
12	takoyaki	43,992
13	fried rice	41,005
14	rice	38,906
15	gratin	36,551
16	sashimi	36,509
17	fried noodle	35,257
18	oden	32,129
19	sashimi bowl	29,478
20	beef bowl	28,974
21	hamburger	24,307
22	cutlet curry	23,401
23	rice ball	22,649
24	croquette	20,854
25	pork cutlet on rice	20,446
26	tempura bowl	20,433
27	tempura	20,281
28	stew	20,169
29	miso soup	20,068
30	chicken rice bowl	19,727
31	taiyaki	19,165
32	sukiyaki	17,985
33	chilled noodle	17,248
34	spicy chili-flavored tofu	16,658
35	egg sunny-side up	15,417
36	croissant	14,165
37	yakitori	13,974
38	omelet	12,859
39	seasoned beef with potatoes	11,953
40	fermented soybeans	11,850
41	rolled omelet	11,536
42	bibimbap	11,148
43	spaghetti meat sauce	9,048
44	simmered pork	8,586
45	spaghetti	8,424
46	steamed egg hotchpotch	8,323
47	eels on rice	7,419
48	pork miso soup	7,366
49	lightly roasted fish	6,914
50	ginger pork saute	6,517
51	udon noodle	5,695
52	egg roll	5,553
53	cabbage roll	5,466
54	fried shrimp	5,244
55	sauteed vegetables	5,210
56	french fries	5,210
57	potato salad	5,034
58	sweet and sour pork	4,969
59	sushi bowl	4,821
60	shrimp with chill source	4,663
61	pilaf	4,483
62	potage	4,356
63	soba noodle	4,147
64	beef noodle	3,434
65	sandwiches	3,289
66	pizza toast	3,247
67	chicken rice	2,877
68	sausage	2,769
69	cold tofu	2,696
70	dried fish	2,694
71	hot dog	2,624
72	steamed meat dumpling	2,526
73	mixed rice	2,331
74	boiled chicken and vegetables	2,076
75	stir-fried beef and peppers	1,855
76	sirloin cutlet	1,507
77	fried chicken	1,456
78	roast chicken	1,296
79	nanbanzuke	1,276
80	roll bread	1,248
81	macaroni salad	1,147
82	boiled fish	774
83	raisin bread	745
84	goya chanpuru	724
85	tempura udon	706
86	kinpira-style sauteed burdock	675
87	chinese soup	571
88	green salad	565
89	Japanese tofu and vegetable chowder	445
90	salmon meuniere	433
91	chip butty	378
92	grilled pacific saury	354
93	tensin noodle	285
94	fried fish	148
95	grilled salmon	123
96	ganmodoki	109
97	vegetable tempura	105
98	sauteed spinach	42
99	grilled eggplant	4
100	teriyaki grilled fish	0
	TOTAL	2,308,988
	(non_food)	13,735,102

Figure 9 shows the ratios of five typical photo genres of geotagged Twitter Photos. From this result, we found no people photos were posted in East Asia, and instead many building and food photos, the total ratio of which were more than 70%, were posted (Figure 5). In North America, the ratio of people and building (Figure 6) were high, which was more than 60%. Regarding South America, people photos are by far the most popular genre (67%) (Figure 7). In Europe, the number of posted photos was the most and the genres were well balanced. In Africa, almost no building, scene and food photos were posted and people photos occupied 70%. In Middle East, although the number of posts were fewer than Europe, all the five genres were balanced as well (Figure 8). In South Asia, more than half of the photos were document photos. This tendency was not observed in other regions. In South-East Asia, in the same way as South America and Africa, people photos are the most and in addition food photos was the second most. Regarding the absolute number of food photos, South-East Asia was the best. Regarding Oceania, the number of the photos was the least among the nine regions. Note that in Oceania, many map photos were posted, and we classified map photos as document photos. That is why the ration of document photos was the best in Oceania.

3.3. Summary of Section

In this experiments, we found the five typical genres which were appropriate for tendency analysis. However, because we discarded small clusters having less than one hundred images in this work, no food and scene photos were left in the photo set of Africa.

From the typical genre distributions, we found the regional tendencies that the ratio of food photos was relatively high in East Asia and East-South Asia, while the ratio of people photos are exceptionally high in South America, South Asia and East-South Asia. In Europe and Middle-East, the typical five genres were well balanced. In addition, almost no people photos are posted in East Asia, and in South Asia half of the posted photos are document photos. In this work, we limited to use only geotagged photos. In general, posting geotagged photos to SNSs means making the current location of the user open to the public. Therefore, we expect that the users in East Asia tend to refrain from posting geotagged people photos to Twitter stronger than normal non-geotagged photos. These are the finding of the analysis in this paper. From these observation, we can estimate that the users in East Asia enjoy posting food photos and the uses in South America, South Asia and East-South Asia like to post people photos without caring privacy issue.

4. WORLD FOOD PHOTO ANALYSIS

In this section, we describe an analysis on Twitter world food photos in order to discover the regional food trends with the same appoarch as the previous section.

Fig. 2. Examples of automatically detected food photos with the proposed DCNN from the Twitter stream. (From the top) ramen noodles, dipping noodles (tsukemen), sushi and omelet.

We use geo-tagged images gathered from the Twitter stream to analyze regional trends of several representative food categories over the world, because food photo mining described in Section 2 is limited to Japanese 100 category foods. This is why we need to classify food and non-food photos again.

The procedure consists of the following four steps as shown in Figure 10:

1. Classify food and non-food photos with a fine-tuned food/non-food classifier for Twitter photos.
2. Extract food-specialized CNN features with a CNN fine-tuned with Triplet loss [20].
3. Perform clustering on food images.

4.1. Classification of food/non-food photos

Food mining described in Section 2 is limited to Japanese 100 category foods, since we used the UECFOOD100 dataset for training of a food classifier. Because the target of this work is world-wide food photo analysis, we use the raw geo-tagged photos gathered from Twitter instead. To select only food images, we need to prepare a food/non-food classifier. We fine-tune the ResNet [21] pre-trained with ImageNet using food image datasets and non-food image datasets. As food image datasets, we use all the images in both the UEC-FOOD100 [22] dataset containing 14,000 images and the Food-101 [23] dataset containing 101,000 images. As non-food image datasets, we use about 120,000 randomly extracted images from the ILSVRC2012 version of the ImageNet 1000-class dataset and about 13,000 non-food images used in the work of Kawano et al.[24]. We discriminate food images from non-food images with this classifier, and discard all the non-food images.

Fig. 3. Examples of misclassified Twitter food photos. Eaten ramen bowl (recognized as "ramen"), unopened instant ramen (ramen), a clam (sushi), ice cream (sushi), an eaten plate (omelet) and curry without cutlet (cutlet curry).

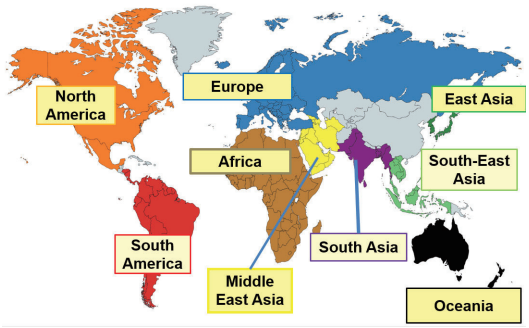

Fig. 4. Nine regions over the world.

4.2. Food image feature extraction

To perform clustering food images, we use CNN features of the pre-trained VGG16 [25]. Since we focus on only food photos, we fine-tune the VGG16 so that it can extract food-specialized features which can discriminate small differences on various kinds of food images. To do that, we use the Triplet network [20] as a method of feature learning. It is known that the Triplet network can improve the image retrieval accuracy of food images [26], which is expect to allow similar kinds of food images to have closer features to each other than the CNN features pre-trained with the ImageNet dataset. The Triplet network [20] is trained by a triplet of a query image, a positive image and a negative image.

For fine-tuning of the VGG16 network for food feature extraction with the triplet loss, we used the images of both UECFOOD100 and FOOD-101 by integrating them into one

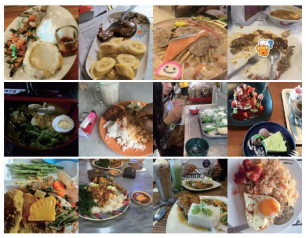

Fig. 5. "Food" in East Asia.

Fig. 6. "Building" in North America.

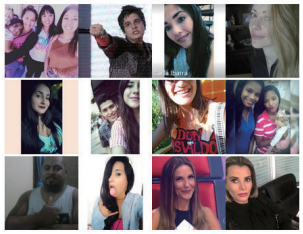

Fig. 7. "People" in South America.

Fig. 8. "Document" in Middle East.

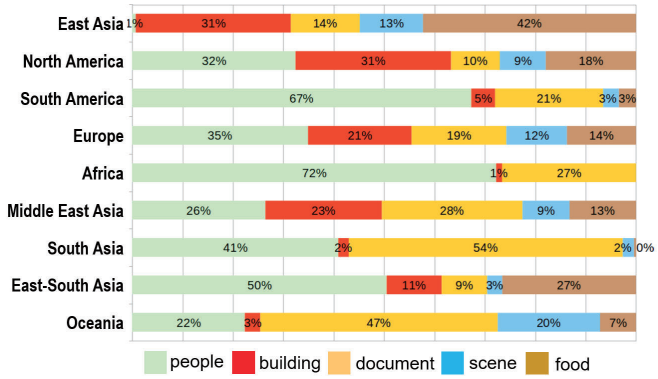

Fig. 9. The ratios of five photo genres in the nine regions over the world.

training image dataset. We extract the 4096-d CNN features from the FC7 of the trained network, normalize them with L2 norm, and then compress them with PCA to 128-d vectors to make large-scale clustering feasible in the same way as the previous section.

4.3. Clustering and analyzing of regional tendency

In the same as the previous work, to analyze Twitter images in the unsupervised way, which means analysis without textual label information, we use a common clustering methods, K-means clustering. Because food CNN features reflect semantic meaning of food images, clustering of food images with food CNN features enables grouping of the food images which are semantically similar to each other [26].

We perform K-means clustering for the food images in each of the pre-defined regions using the food-specialized CNN features. After clustering, we obtain clusters of the

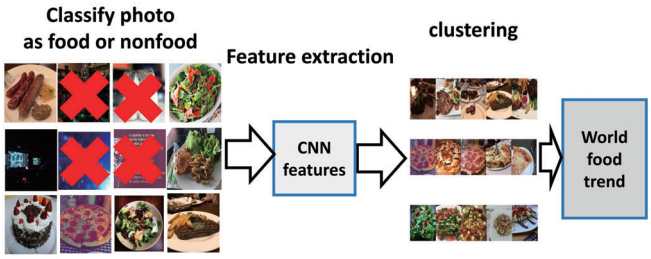

Fig. 10. The proposed procedure for analyzing regional food trends.

Table 4. Image statistics.

region	East Asia	South Asia	South-East Asia	North America	South America
all	488,609	79,179	609,671	786,093	410,086
food	69,826	999	59,459	23,867	8,211
region	Europe	Africa	Oceania	Middle East	**TOTAL**
all	791,095	150,550	34,543	434,790	**3,784,616**
food	14,572	1,229	914	14,115	**193,192**

images which were semantically similar to each other. To analyze the tendency of posted food photos, we classify the obtained photo clusters into one of the pre-selected representative food categories. As the representative food categories, we use 17 kinds of foods such as "meat", "noodle", "rice" and "bread". These categories are decided based on the observation of clustering results of each of the regions. After representative category classification of clusters, we compare the food distributions of the images in each region regarding all the region to clarify the difference of regional tendency of Tweet photos. When showing the photos within each cluster, we sort photos within each cluster by a similarity-based image ranking method, VisualRank [27].

4.4. Analysis on World Twitter Food Photo

In this work, we used the log of the Tweets containing both photos and geo-tags we collected in 2016 for whole a year. With the food/non-food classifier, we selected 220,000 food images from 3.78 million geotagged Twitter images contained in the Twitter log. After that, duplicate images were removed based on simple color histogram image features. Finally, we obtained about 190,000 food images. The number of all images and food images for each region are shown in Table 4. In the analysis, we used food images in six regions, East Asia, South-East Asia, North America, South America, Europe and the Middle East excluding three regions, South Asia, Africa and Oceania, because the number of geotagged food images in the three regions were around 1,000 which were not enough as shown in the table.

We performed K-means clustering for each of the seven region with the CNN features extracted from the food image, and classified the obtained clusters into the representative food categories. The food categories assigned to the cluster

Table 5. List of the representative food categories.

meat salad egg	noodles fried-food curry	sweets seafood flour	rice soup Chinese cuisine	bread fastfood cheese	beverage stir-fried

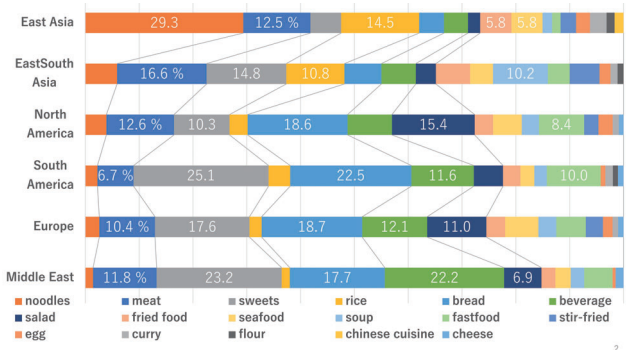

Fig. 11. Food distribution on the six regions. White letters indicate the percentage of the top 5 categories on each region.

Fig. 12. Examples of "noodles" images in East Asia.

Fig. 13. Examples of "soup" images in South-East Asia.

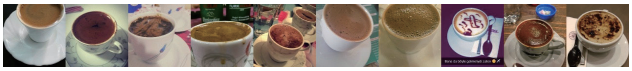

Fig. 14. Examples of region-specific "beverages" in Middle East.

are 17 categories shown in Table 5. These categories are decided based on the observation of clustering results of each of the regions. The datasets used for training the food/non-food classifier, UECFOOD100 [22] includes images such as "ramen" and "pilaf", and FOOD-101 includes images such as "apple_pie" and "french_onion_soup." We selected 17 representative food categories so that they roughly covered all the categories of both the datasets.

First, we analyzed the ratio of the food categories on each region. The ratios on each region are shown in Figure 11 where the percentages of the top five food categories among 17 categories are shown in bold.

In East Asia, the food categories such as "noodles" and "rice", which are relatively rare in the other regions, are included at the top. As the possible reason for the high proportion of noodles to 29.3%, there are many images of "ramen" (which are very popular noodles in Japan.) as shown in Figure 12, and in addition to "ramen", various kinds of "noodle" images such as rice noodles and buckwheat noodles are seen in East Asia. This is a salient characteristics of East Asia.

In South-East Asia, the soup is ranked at the top, and many dishes made of vegetables and meats in a soup like Figure 13 were seen.

In North America, when comparing trends in Asia, the categories for bread and salad tended to be more. On the other hand, when compared with South America and Europe, it turned out that 4 items of the top 5 items of the categories are the same, and two regions had similar to each other regarding food tendency.

In the Middle East, the top five food categories are the same as Europe. However, in the region there are many brown coffee as shown in Figure 14, which is a unique type of coffee to Middle East.

4.5. Summary of Section

In this section, we analyzed regional food tendency on only the geo-tagged food images without text data. As a result of the analysis, we could see the unique food tendency on each region and similar food image tendency with other regions.

In this work, we were unable to analyze regional trends in the three regions, South Asia, Africa, and Oceania. This was because there were extremely few food images posted on Twitter in the three areas. For future work, we plan to integrate image data from other photo SNSs such as Instagram and Weibo with Twitter images for more comprehensive analysis. Since we have been collecting Twitter photo logs from 2011 continuously, we also plan to analyze temporal transition of Twitter photo trends for about eight years.

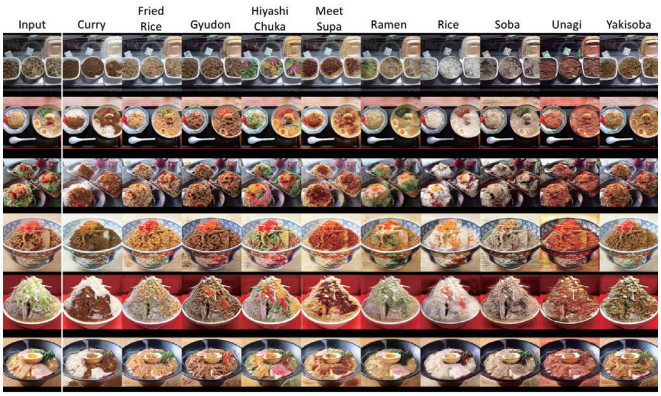

Fig. 15. The leftmost images are input images, and the other ones are generated regarding each of the ten categories.

5. APPLICATIONS OF TWITTER FOOD PHOTO

In this section, as applications of the food photo database we created by image gathering from Twitter, we briefly introduce food image translation by StarGAN [28], and food image generation by conditional StyleGAN [14], both of which are variations of generative adversarial networks (GAN). To train GANs effectively, a large number of images are required. Twitter photo mining matches well to this requirement, since we can gather a large number of photos from Twitter easily.

Fig. 16. The leftmost images are the input images. The remaining six images are separated into two blocks. The left blocks show the results of "beef bowl" and the right blocks show the results of "curry rice." In each block, from left to right, we show the generated images trained with with 10,000 images, 100,000 images, and 230,000 images, respectively.

For both works, we used more than 10,000 Twitter food photos per category.

5.1. Food Image Translation

We applied StarGAN [28] which is a conditional version of CycleGAN [29] into our large-scale food dataset, and realized food image translation among ten kinds of Japanese foods [9, 10]. We can translate an input food image into any of the ten kinds of food images as shown in Figure 15. For example, in the fourth row, a photo of beef rice bowl was converted into curry, fried rice, meat spaghetti, ramen and so on. Only the food region was changed.

We show the food image transformation results, when we used a smaller dataset for training the model. Here, we prepare the following three types of subsets of the dataset.

(1) 1000 image per category: 10,000 images.

(2) 10000 image per category: 100,000 images.

(3) All the images: 230,053 images.

In Figure 16, we show the results obtained from the model trained with different number of images. The leftmost images are the input images, and the remaining six images are generated images. The transformed images are separated into two blocks by the food categories used for the conditional vector. In each block, we used 10,000 images for the first column, 100,000 images in the second column, and 230,000 images in the third column, for training. The generated image quality becomes better as the number of training images increases. Although we obtained acceptable results by the model trained with a small training set, the details are not reconstructed.

The GAN-based food translation was imported to the virtual reality (VR) domain. We achieved real-time food translation on head mount displays (HDM) [11, 12], which was difficult to implement by conventional computer graphics methods.

Fig. 17. Examples of generated images with (a) single food conditions (b) mixing conditions of two foods (c) mixing conditions of three foods.

5.2. Food Image Generation

In the end of this paper, we like to introduce food image generation by conditional StyleGAN [14] briefly. StyleGAN is the state-of-the-art in terms of image quality and the ability of style manipulation. We extended this by adding conditional inputs which correspond to food categories of generated food images.

Figure 17 shows some examples of generated foods with single conditions (at the first row), mixing conditions of two foods (at second row) and mixing conditions of three foods (at third row). By mixing food category conditions, we can generate unseen mixed foods of multiple categories. This is an interesting characteristic of conditional GANs. We might create novel delicious-looking foods with the GAN trained with the large-scale Twitter food photo database.

6. CONCLUSIONS

In this paper, we introduced our works on Twitter photos especially related to food domain. We explained the current food photo mining system at first. Next, we introduced Twitter world photo analysis [7] and world food photo analysis [8]. As applications of Twitter food photo mining, we described food photo translation [9, 10], and generation [14].

For future works, we think that Twitter food mining with GAN is one of the interesting research directions. By training from Twitter food photos over the world, we might be able to realize the GAN which can generate any kinds of regional foods.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 15H05915, 17H01745, 17H06100 and 19H04929.

7. REFERENCES

[1] K. Yanai, "World seer: A realtime geo-tweet photo mapping system," in *Proc. of ACM International Conference on Multimedia Retrieval*, 2012.

[2] K. Yanai and Y. Kawano, "Real-time food image mining and analysis from the twitter stream," in *Proc. of Pacific-Rim Conference on Multimedia (PCM)*, 2014.

[3] Y. Nakaji and K. Yanai, "Visualization of real world events with geotagged tweet photos," in *Proc. of IEEE ICME Workshop on Social Media Computing (SMC)*, 2012.

[4] T. Kaneko and K. Yanai, "Visual event mining from geo-tweet photos," in *Proc. of IEEE ICME Workshop on Social Multimedia Research (SMMR)*, 2013.

[5] T. Kaneko and K. Yanai, "Event photo mining from twitter using keyword bursts and image clustering," *Neurocomputing*, vol. 172, pp. 143–158, 2016.

[6] S. Matsuo, W. Shimoda, and K. Yanai, "Twitter photo geo-localization using both textual and visual features," in *Proc. of International Conference on Multimedia Big Data (BIGMM)*, 2017.

[7] T. Nagano, T. Ege, W. Shimoda, and K. Yanai, "A large-scale analysis of regional tendency of twitter photos using only image features," in *Proc. of IEEE International Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2019.

[8] K. Okamoto and K. Yanai, "Analyzing regional food trends with geo-tagged twitter food photos," in *Proc. of International Conference on Content-Based Multimedia Indexing (CBMI)*, 2019.

[9] R. Tanno, D. Horita, W. Shimoda, and K. Yanai, "Magical rice bowl: Real-time food category changer," in *Proc. of ACM International Conference Multimedia*, 2018.

[10] D. Horita, R. Tanno, W. Shimoda, and K. Yanai, "Food category transfer with conditional cycle gan and a large-scale food image dataset," in *Proc. of International Workshop on Multimedia Assisted Dietary Management (MADIMA)*, 2018.

[11] S. Naritmo, R. Tanno, T. Ege, and A. K. Yanai, "Cnn-based food transformation on hololens," in *Proc. of International Workshop on Interface and Experience Design with AI for VR/AR (DAIVAR)*, 2018.

[12] K. Nakano, D. Horita, N. Sakata, K. Kiyokawa, K. Yanai, and T. Narumi, "Enchanting your noodles: A gustatory manipulation interface by using gan-based real-time food-to-food translation," in *Proc. of IEEE Virtual Reality*, 2019.

[13] Y. Ito, W. Shimoda, and K. Yanai, "Food image generation using a large amount of food images with conditional gan: Ramengan and recipegan," in *Proc. of International Workshop on Multimedia Assisted Dietary Management (MADIMA)*, 2018.

[14] D. Horita, W. Shimoda, and K. Yanai, "Unseen food creation by mixing existing food images with conditional stylegan," in *Proc. of International Workshop on Multimedia Assisted Dietary Management (MADIMA)*, 2019.

[15] K. Yanai and Y. Kawano, "Food image recognition using deep convolutional network with pre-training and fine-tuining," in *Proc. of ICME Workshop on Multimedia for Cooking and eating Activities (CEA)*, 2015.

[16] Y. Matsuda and K. Yanai, "Multiple-food recognition considering co-occurrence employing manifold ranking," in *Proc. of IAPR International Conference on Pattern Recognition*, 2012.

[17] A Krizhevsky, I Sutskever, and G E Hinton, "Imagenet classification with deep convolutional neural networks.," in *Advances in Neural Information Processing Systems*, 2012.

[18] G. Huang, Z. Liu, L. V. Der, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2017.

[19] N. Martinel, G. L. Foresti, and C. Micheloni, "Wide-slice residual networks for food recognition," in *Proc. of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018.

[20] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, B. Philbin, J. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *Proc.of IEEE Computer Vision and Pattern Recognition*, 2014.

[21] K. He, X. Zhang, and J Ren, S.and Sun, "Deep residual learning for image recognition," in *Proc.of IEEE Computer Vision and Pattern Recognition*, 2015.

[22] Y. Matsuda, H. Hoashi, and K. Yanai, "Recognition of multiple-food images by detecting candidate regions," in *ICME*, 2012.

[23] B. Lukas, G. Matthieu, and V.G. Luc, "Food-101 – mining discriminative components with random forests," in *Proc.of European Conference on Computer Vision*, 2014.

[24] Y. Kawano and K. Yanai, "Automatic expansion of a food image dataset leveraging existing categories with domain adaptation," in *Proc.of ECCV Workshop on Transferring and Adapting Source Knowledge in Computer Vision (TASK-CV)*, 2014.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[26] W. Shimoda and K. Yanai, "Learning food image similarity for food image retrieval," in *Proc.of IEEE BIG Multimedia (BIGMM)*, 2017.

[27] Y. Jing and S. Baluja, "VisualRank: Applying pagerank to large-scale image search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 1870–1890, 2008.

[28] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-imag e translation," in *Proc.of IEEE Computer Vision and Pattern Recognition*, 2018.

[29] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc.of IEEE International Conference on Computer Vision*, 2017.