Gaze-inspired Learning for Estimating the Attractiveness of a Food Photo

Akinori Sato
Graduate School of Infomatics
Nagoya University
Nagoya, Japan
satoa@murase.is.i.nagoya-u.ac.jp

Takatsugu Hirayama
Institutes of Innovation for Future Society
Nagoya University
Nagoya, Japan
takatsugu.hirayama@nagoya-u.jp

Keisuke Doman
School of Engineering
Chukyo University
Toyota, Japan
kdoman@sist.chukyo-u.ac.jp

Yasutomo Kawanishi
Graduate School of Infomatics
Nagoya University
Nagoya, Japan
kawanishi@i.nagoya-u.ac.jp

Ichiro Ide
Graduate School of Infomatics
Nagoya University
Nagoya, Japan
ide@i.nagoya-u.ac.jp

Daisuke Deguchi
Information Strategy Office
Nagoya University
Nagoya, Japan
ddeguchi@nagoya-u.jp

Hiroshi Murase
Graduate School of Infomatics
Nagoya University
Nagoya, Japan
murase@i.nagoya-u.ac.jp

***Abstract*—The number of food photos posted to the Web has been increasing. Most of the users prefer to post delicious-looking food photos. They, however, do not always look delicious. A previous work proposed a method for estimating the attractiveness of food photos, that is, the degree of how much a food photo looks delicious, as an assistive technology for taking a delicious-looking food photo. This method extracted image features from the entire food photo to evaluate the impression. In our work, we conduct a preference experiment where subjects are asked to compare a pair of food photos and measure their gaze. The proposed method extracts image features from local regions selected based on the gaze information and estimates the attractiveness of a food photo by learning regression parameters. Experimental results showed the effectiveness of extracting image features from outside the gaze regions rather than inside them.**

***Keywords*-Food photography; attractiveness; gaze**

I. Introduction

The number of food photos posted to the Web has been increasing with the spread of Social Networking Services and cooking recipe portal sites. Most of the users prefer to attract other users' interests by posting delicious-looking food photos. The food photos, however, do not always look delicious since they are taken by an amateur photographer. Fig. 1(a) would look delicious than Fig. 1(b) in terms of camera angle and its photographic framing, although these two photos actually took the same food. In order to take a delicious-looking food photo, we need to select an appropriate shooting approach. Thus, it would be useful to realize a system that can help us to take attractive food photos or a system for selecting the most attractive one from a list of food photos. Here, it is necessary to quantitatively analyze the attractiveness of a food photo. We define the attractiveness as the degree of how much a food photo looks delicious.

For evaluating the aesthetics of general photos, Nishiyama et al. proposed a method to classify whether the aesthetic quality is good or not considering color harmony and color variations in local regions [1]. Tian et al. also proposed a method to construct binary aesthetics classification model for each image query using deep convolutional neural networks (DCNNs) [2]. However, since these methods are specialized to binary classification of the aesthetics quality of photos, it is difficult to rank and select the best one from a list of photos.

(a) Attractive framing (b) Non-attractive framing

Figure 1: Photographic framing of a food.

For photography assistance, Kakimori et al. developed a system that shows a user the guideline for arranging dishes in photographic framing [3]. Although the system can recommend an attractive dish arrangement, the system neither recommends the best camera angle for each food nor evaluates the attractiveness of food photos. Meanwhile, Takahashi et al. proposed a method for estimating the attractiveness of food photos based on several kinds of image features, and constructed an image dataset where the attractiveness is assigned to each photo through preference experiments by subjects [4]. This method extracts a Deep Convolutional Activation Feature (DeCAF) [5] in addition to several color and shape features to evaluate the impression of the entire food photo and the appearance of the main ingredient. They confirmed the effectiveness of the method using the dataset with ten food categories. Although this

978-1-5386-6857-3/18/$31.00 ©2018 IEEE
DOI 10.1109/ISM.2018.00015

work assumed that it is important to extract useful features from the region of the main ingredient, other useful regions for estimation seem to exist. Here, in our work, we focus on a region where the viewers generally put their gaze on to find more useful regions. This is because we assumed that the viewer's gaze is a clue to search for regions including important features for estimation. By extracting image features from such regions, we expect to improve the accuracy of estimating the attractiveness of food photos.

As related work, Shimojo et al. performed an experiment where subjects selected one out of two facial images arranged horizontally based on preference, i.e. pairwise comparison, and showed that the gaze distribution of the subjects was biased towards the image selected [6]. Sugano et al., by applying this finding, proposed a method to estimate the preference of a natural image using the viewer's gaze information and confirmed a strong relationship between gaze behavior and preference of an image [7]. Matsumoto et al. confirmed the improvement in the performance of a pedestrian's gender classifier by emphasizing the image features of the region around the viewer's gaze point [8].

Therefore, in order to improve the accuracy of estimating the attractiveness, we decided to design the features based on the findings from gaze analysis. Since the attractiveness value defined in the previous work [4] was determined by pairwise comparison of food photos, we also analyze the gaze during pairwise comparison. We therefore conduct pairwise comparison experiments to measure the gaze and analyze gaze points to select regions for extracting useful image features. In this paper, we propose a method to estimate the attractiveness of food photos using image features from such regions.

In short, the contribution of this paper is the improvement of the attractiveness estimation accuracy by using image features extracted from the local regions selected based on gaze information.

This paper is organized as follows. Section II describes the details of the proposed method. Next, dataset construction through subjective experiments is described in Section III. Then, evaluation of the proposed method is reported in Section IV. Finally, Section V concludes this paper.

II. Proposed Method

Fig. 2 shows the process-flow of the proposed method composed of three steps: dataset construction, training, and estimation. In the dataset construction step, we take dishes from multiple directions. Then, experimental subjects compare a pair of food photos taken from different view points according to their preference. The gaze information is acquired as a time series of the coordinates of gaze points on each food photo. The proposed method selects regions based on the accumulation of gaze information for extracting image features. The training step extracts image features from the selected regions and constructs an attractiveness estimator by learning the relationship between the image features and the attractiveness values of a food photo. The estimation step estimates the attractiveness of an input food photo whose attractiveness is unknown using the constructed attractiveness estimator. Note that the gaze information is not given to the input food photo but we assume that the view point is given as a known parameter.

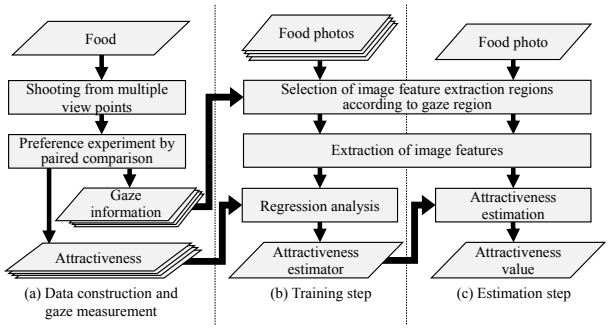


Figure 2: Process-flow of the proposed method.

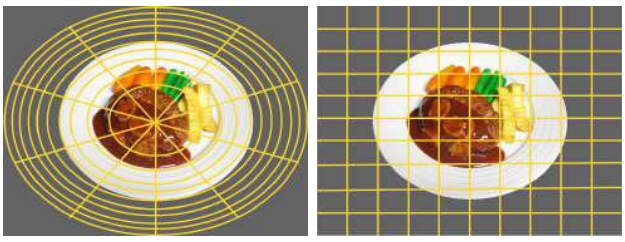
The following sections describe the procedure of the training and the estimation steps. The data construction step is described in Section III.

A. Training Step: Construction of an Attractiveness Estimator

The training step constructs an attractiveness estimator as shown in Fig. 2(b). First, for each food photo with an attractiveness value, a dish region including all the ingredients is cropped by GrabCut [9]. Next, the gaze information is analyzed, and the gaze regions are detected according to the cumulative fixation time while fixing the gaze on each of the divided regions shown in Fig. 3. Then, regions for extracting image features are selected by thresholding the cumulative time. Finally, an attractiveness estimator is constructed using a regression framework. We use the Random Regression Forests [10] for the purpose. Here, the objective variable is the attractiveness value of a food photo, and the explanatory variables are the image features extracted from the selected image feature extraction regions.

The following sections describe how to extract color feature $\boldsymbol{C}$ and shape feature $\boldsymbol{E}$ in detail, which were introduced in the previous work [4]. Although the proposed method concatenates $\boldsymbol{C}$ and $\boldsymbol{E}$, we can also use them individually to see which feature is effective for the estimation.

1) Color Feature: It is known that color harmony is important when judging the aesthetics of photography [1]. Based on this knowledge, we focus on the hue of ingredients and the color harmony in a dish. Thus, the proposed method measures the color difference in the CIELAB color space, which is designed to approximate human visual perception in color difference.



(a) Division for the color feature extraction (b) Division for the shape feature extraction

Figure 3: Region division for extracting image features.

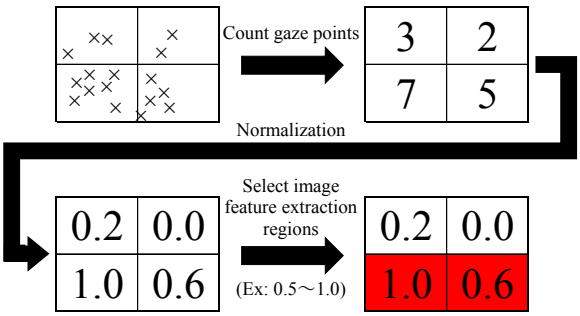


Figure 4: Example of selecting image feature extraction regions.

First, the most frequent color (L_m, a_m, b_m) in the CIELAB color space is calculated from the entire dish region. Each of the color channels here is quantized into eight levels $(1 \leq L_m, a_m, b_m \leq 8)$ to reduce the number of dimensions of the feature vector. Next, the input image is divided into 100 radial local regions as shown in Fig. 3(a), and the most frequent color (L_i, a_i, b_i) and its frequency F_i are calculated in each local region. Here, i indicates the index of each block, $1 \leq i \leq 100$, and $1 \leq L_i, a_i, b_i \leq 8$. Next, the color difference C_i is calculated as

$$C_i = F_i \sqrt{(L_m - L_i)^2 + (a_m - a_i)^2 + (b_m - b_i)^2}. \quad (1)$$

Finally, a 100-dimensional vector $\boldsymbol{C} = (C_1, C_2, \ldots, C_{100})$ is obtained as a color feature.

2) Shape Feature: It is known that the shape and the arrangement of ingredients affect the visual appearance of food photos [11]. Based on this knowledge, we focus on the geometric pattern of a food. Thus, the proposed method extracts gradient features of the intensity of an image.

First, an input image is divided into 10×10 local regions as shown in Fig. 3(b). Next, the maximum edge strength e_j and the gradient orientation n_j from each local region are calculated and multiplied. Here, j indicates the index of each local region, and $1 \leq j \leq 100$. We ignore the features extracted from around the edge of the dish (5 pixels range) because we do not focus on its shape. The gradient orientation from each local region is quantized into 36 levels to reduce the number of dimensions of the feature vector. Finally, a 100-dimensional vector $\boldsymbol{E} = (e_1 n_1, e_2 n_2, \ldots, e_{100} n_{100})$ is obtained.

3) Selection of Image Feature Extraction Regions by Analyzing Gaze: Regions for extracting the image feature used for the estimation are selected based on the gaze information. First, gaze regions are detected according to the cumulative fixation time while fixing the gaze on each local region defined in Section II-A1 and Section II-A2. Here, the fixation is a state in which the gaze point remains at a certain position continuously. We defined that two consecutive gaze points whose gaze movement angular velocity is less than 30 degrees per second are in the fixation state. The motion speed of gaze points is calculated from the distance between two consecutive gaze points, assuming that the gaze points are measured at fixed time intervals. As preprocessing, from the gaze points obtained during the subjective experiment described in Section III, the fixations are extracted in the dish region. Then, normalization is performed over all the local regions so that the minimum and the maximum of the cumulative fixation time becomes 0 and 1, respectively. Finally, by setting a range of the cumulative fixation time, the image feature extraction regions are selected. Note that even if the gaze information is not available, if a view point is given, the image feature extraction regions can be defined by using rotated gaze points on other images taken from different view points. Fig. 4 shows an example of selecting image feature extraction regions by analyzing gaze information.

B. Estimation Step: Attractiveness Estimation

The estimation step estimates the attractiveness of food photos by the procedure shown in Fig. 2(c). Note that although gaze is not measured in the estimation step, we assume that the view point is given as a known parameter. Thus the image feature extraction regions are identified by rotating gaze points on other images shot from different view points. First, GrabCut [9] is applied to the input image to extract the dish region containing all the ingredients. Next, from the dish region, image features are extracted only from the image feature extraction regions selected through the analysis described in Section II-A3. Finally, we estimate the attractiveness using the attractiveness estimator described in Section II-A.

III. Dataset Construction

The proposed method selects image feature extraction regions through the analysis of gaze measured by conducting a preference experiment by subjects. In this section, this experiment for constructing the dataset is described in detail.

(a) Sashimi (b) Curry and rice (c) Eel rice-bowl (d) Beef stew (e) Hamburger steak

(f) Tempura rice-bowl (g) Fried pork rice-bowl (h) Tuna rice-bowl (i) Cheese burger (j) Fish burger

Figure 5: Food categories in "NUFOOD 360×10" [4].

A. *Food Photo Dataset with Attractiveness*

We used "NU FOOD 360×10"[1], a food photo dataset with attractiveness constructed in [4]. The dataset includes ten food photo groups taken from 36 view points. These photos are used for the preference experiment. Each photo is already assigned an attractiveness value calculated by Thurstone's pairwise comparison method [12].

Here, we also used each food photo as an image presented to subjects in the gaze measurement experiment described later in Section III-B and as an input image in the evaluation experiment reported in Section IV. Also, the attractiveness assigned to each food photo is used as the target value of the attractiveness estimator in the evaluation experiment later reported in Section IV. Details of this dataset is introduced below.

1) Food Categories: The dataset includes ten food categories shown in Fig. 5, namely, Sashimi, Curry and rice, Eel rice-bowl, Beef stew, Hamburger steak, Tempura rice-bowl, Fried pork rice-bowl, Tuna rice-bowl, Cheese burger, and Fish burger. These food categories were selected considering the difference in color, shape, and solidity. Plastic food samples were used instead of real ones considering both convenience and reproducibility.

2) Photographing Method: Food photos were taken from three elevation angles: 30, 60, and 90 degrees. Also, an arbitrary rotation angle was set as 0 degrees, and the food photos were taken from 0 to 330 degrees with a step of 30 degrees in clockwise direction around the center of the dish. As a result, 36 food photos were obtained for each food category. The width of each image was 720 pixels.

3) Determination of Attractiveness Values by Paired Comparison: An attractiveness value was assigned to each food photo by Thurstone's pairwise comparison method [12]. This method is one of sensory tests, and scales the sensory values of samples based on a number of paired comparison results. 360 ($= {}_{36}\mathrm{C}_2$) photo pairs were shown one by one to human subjects, who were asked to respond which photo looked more delicious by selecting one of the buttons: "Left," "Right," or "Difficult to say." The subjects were 28 students in their twenties. As a result, three or four responses were obtained for each photo pair and 2,150 responses in total for each food category. Finally, an attractiveness value was calculated from paired comparison results, and normalized into the range of [0,1].

B. *Gaze Measurement during Pairwise Comparison*

In this section, we describe the gaze measurement while conducting a preference experiment by subjects using the dataset introduced in Section III-A.

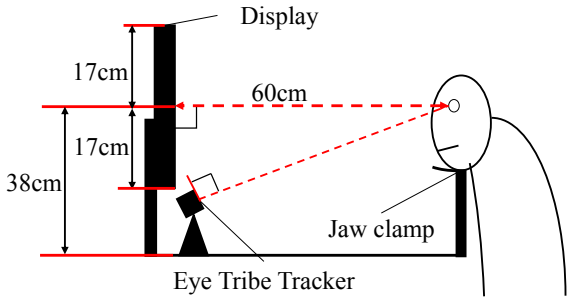
1) Environment: In order to measure a subject's gaze, a display, an eye tracking device, and a jaw clamp were prepared. The experimental setting for the gaze measurement is shown in Fig. 6. We used Eye Tribe Tracker[2] as an eye tracking device, and set its sampling frequency to 30 Hz. The size of the display was 27 inches, and the resolution was $1,920 \times 1,080$ pixels. We fixed the distance between the display and the jaw clamp to about 60 cm, and adjusted the eye positions of the subjects to the height of the center of the display as shown in Fig. 6(b). Note that in this setting, one degree of the central visual field of each subject corresponds to a circle with a diameter of about 34 pixels on the display.

2) Dataset: For the food categories presented to the subjects, we chose Hamburger steak and Curry and rice with different appearances from the data set introduced in Section III-A. In order to reduce the number of photo pairs for a subject to compare, we selected ten photos with attractiveness values over 0.5 shown in Fig. 7, which seems to be relatively difficult to estimate the attractiveness among

[1]Available from http://www.murase.is.i.nagoya-u.ac.jp/nufood/

[2]The Eye Tribe Aps, "Eye Tribe Tracker," http://theeyetribe.com/

(a) Experiment scene



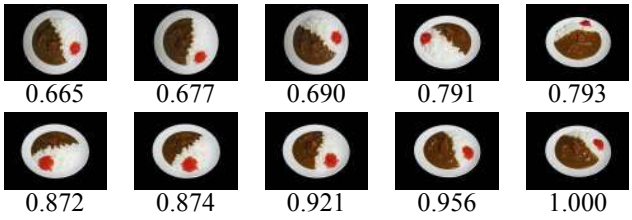
(b) Side view

Figure 6: Experiment setting for the gaze measurement.

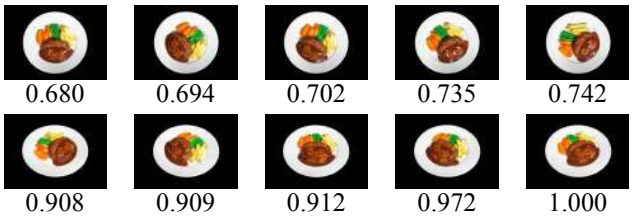
the 36 photos of each category. We generated 90 ($=_{10}\mathrm{P}_2$) photo pairs for each category.

3) Method: First, to fix a subject's gaze to a specific position, a cross is shown in the center of the screen. The subject is instructed to press the Enter key on the keyboard when gazing at the center of the image. Immediately after that, a photo pair generated in Section III-B1 is presented. Similarly to the preference experiment introduced in Section III-A3, the subject is asked to respond which photo seemed more delicious by pressing one of the buttons: "Left," "Right," or "Difficult to say." Since the photo pair was displayed as large as possible on the full screen, it was easy for the subjects to see the details. The gaze points are measured from when a photo pair is displayed until when the button is pressed. After that, the cross to fix the gaze is presented again. This task was conducted for the 90 image pairs for each category by nine students in their twenties.

4) Integration of Gaze Information: In the above experiment, the gaze was measured only for ten out of 36 food photos in each food category. In the evaluation experiment conducted in Section IV, we will estimate the attractiveness for all the 36 photos from different view points. Therefore,



(a) Curry and rice



(b) Hamburger steak

Figure 7: Photos used for gaze measurement. Numbers below the image indicate attractiveness values.

we synthesized gaze information for the other 26 photos as follows. First, one of the 36 photos which were taken from rotation angle α and elevation angle β is defined as a base image which was taken from rotation angle α' and elevation angle β'. Next, the gaze points $(x_g(t), y_g(t))$ for each of the ten photos are rotated so that $(x'_g(t), y'_g(t))$ match the base image as

$$\begin{aligned}
&\begin{pmatrix} x'_g \\ y'_g \\ 1 \end{pmatrix} = A'RA \begin{pmatrix} x_g \\ y_g \\ 1 \end{pmatrix} \\
&A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \mathrm{cosec}\beta & 0 \\ 0 & 0 & 1 \end{pmatrix}, A' = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \sin\beta' & 0 \\ 0 & 0 & 1 \end{pmatrix} \\
&R = \begin{pmatrix} \cos\bar{\alpha} & \sin\bar{\alpha} & X_c - X_c\cos\bar{\alpha} - Y_c\sin\bar{\alpha} \\ -\sin\bar{\alpha} & \cos\bar{\alpha} & Y_c + X_c\sin\bar{\alpha} - Y_c\cos\bar{\alpha} \\ 0 & 0 & 1 \end{pmatrix} \\
&\bar{\alpha} = \alpha' - \alpha.
\end{aligned} \tag{2}$$

Here, (X_c, Y_c) is the center point of a food photo, A is a matrix that rotates the elevation angle from β to 90 degrees, R is a matrix that rotates the rotation angle from α to α', and A' is a matrix that rotates the elevation angle from 90 degrees to β'. Note that formula (2) is represented in the homogeneous coordinate system. Then, the gaze points are integrated by superimposing them onto the base image. Finally, by projecting the integrated gaze points on each photo from each view point using formula (2), pseudo gaze information for the 36 photos is synthesized. Fig. 8 shows an example of this procedure.

5) Results: Figs. 9 and 10 show heat maps based on cumulative fixation time calculated from the integrated gaze

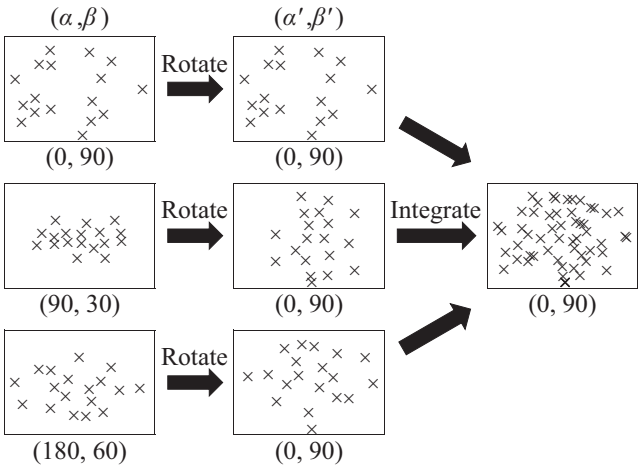


Figure 8: Example of integration of gaze information. Crosses indicate gaze points. Numbers in parentheses indicate (elevation angle, rotation angle).

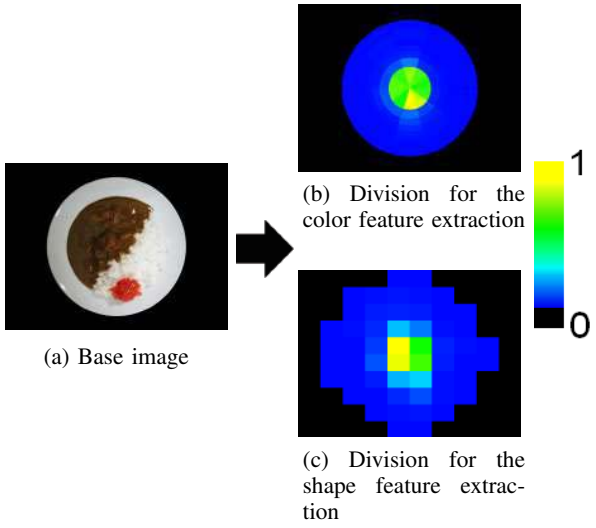


Figure 9: Heat map of cumulative fixation time for Curry and rice.

information by the method described in Section II-A3.

For both dishes, we can see that the cumulative fixation time in the local regions near the center is longer. In other words, regardless of food category or angle, it seems that people spend longer time to look at the center compared to the surroundings.

IV. Evaluation Experiments

We evaluated the effectiveness of the proposed method through experiments.

A. Method

We took a leave-one-out scheme with the dataset described in Section III-A for training and evaluating the

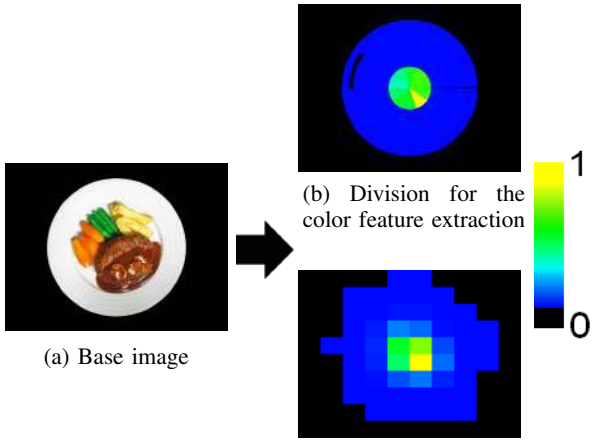


Figure 10: Heat map of cumulative fixation time for Hamburger steak.

proposed attractiveness estimator.

Regarding the selection of the image feature extraction regions, the threshold of cumulative fixation time in each local region was set to 0.1 seconds in order to set the area ratio of the image feature extraction regions and the other regions close to 50%. Here, we regard the regions with a cumulative fixation time longer than 0.1 seconds as the gaze regions.

We compared the estimation accuracy of the proposed method with that of the following methods.

- **Aesthetic**: An aesthetic evaluation method for general photos based on [2], which constructs a deep learning model for each query image.
- **Previous**: A previous method based on [4], which extracts image features from the entire dish region without considering gaze information.
- **Gaussian**: The same method as the proposed method except that the fixation point distribution is assumed to be Gaussian distribution with μ being a center point of a food photo and 3σ being half of each axis of a dish region.

For training of Random Regression Forest, we used `RandomForestRegressor` in the scikit-learn library [13], with parameters `random_state` $= 2$ and `n_estimators` $= 150$. As pre-processing, each feature was normalized to [0,1]. For each method, we calculated the Mean Absolute Error (MAE) between the estimated values and the target values for the attractiveness of food photos as estimation error.

B. Results

Experimental results when combining two image features are summarized in Table I. The MAE was minimized when

Table I: Mean Absolute Error (MAE) when combining two image features. Underlined values indicate equal to or less than the error in [4], and bold letters indicate the minimum value.

		Color		
		Entire [4]	Inside (over 0.1)	Outside (less than 0.1)
Shape	Entire [4]	0.121	0.122	0.114
	Inside (over 0.1)	0.165	0.131	0.131
	Outside (less than 0.1)	0.120	0.124	**0.113**

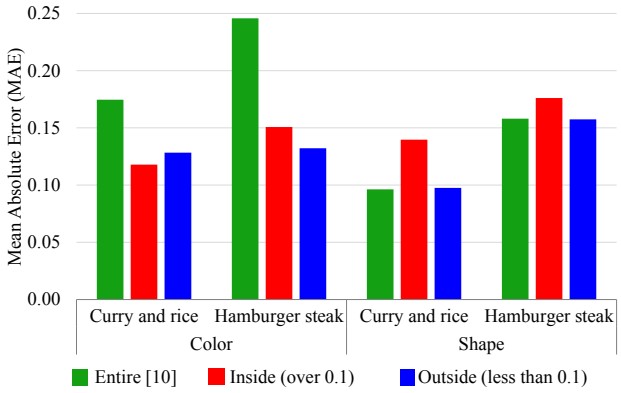


Figure 11: Mean Absolute Error (MAE) when using only one of the image features.

both image features were extracted from outside the gaze regions. However, when extracting either image feature from the gaze regions, the estimation error was worse compared to the previous method. Therefore, we can say that it is not effective to extract image features from the gaze regions.

To show whether the color feature or the shape feature was effective for estimation, the results for each dish when using only one of the features are shown in Fig. 11. In the case of only using the color feature, regardless of whether extracting from inside or outside the gaze regions, the estimation error decreased compared to the previous method. On the other hand, in the case of only using the shape feature, when extracting from the gaze regions, the estimation error increased compared to the previous method, and when extracting from outside the gaze regions, there was almost no difference. Therefore, we can say that it is effective to select regions of the color feature extraction but not the shape feature extraction.

The results are summarized in Table II. Note that Gaussian method and the Proposed method extracts image features from outside the gaze regions. The MAE of the proposed method was minimized in all categories. Therefore, we can say that the proposed method which extracts image features from local regions selected based on the gaze information is effective.

Table II: Comparison of Mean Absolute Error (MAE) in each method. Bold letters indicate the minimum value in each category. Gaussian and Proposed extracts image features from outside the gaze regions.

Method	Mean Absolute Error		
	Curry and rice	Hamburger steak	Average
Aesthetic [2]	0.214	0.258	0.236
Previous [4]	0.093	0.149	0.121
Gaussian	0.094	0.144	0.119
Proposed	**0.085**	**0.142**	**0.113**

C. Discussion

The estimation error decreased by using the image features extracted from outside the gaze regions, whereas the estimation error increased by the extraction from the gaze regions. When rating the attractiveness of a photo, it seems that humans positively use not only information inside the gaze regions but also information outside them, which is different from the results in [8]. Since the subjects in [8] mainly looked at body parts to distinguish the gender difference of a pedestrian such as the head and the chest, the differences in image features between gaze regions seemed to have been large. On the other hand, in our work, regions close to the photo center had higher cumulative fixation time as shown in Figs. 9 and 10. Since the food photos used in our work have only variation of view point and their center corresponds to the center of a dish region, the closer the regions are to the center of a photo, the smaller the differences are in image features between them. Accordingly, the differences in image features between gaze regions seemed to be small. Therefore, we consider that it was difficult to estimate the attractiveness of food photos only by extracting image features from the gaze regions.

The proposed method based on the fixation point distribution reduced the estimation error than the method based on Gaussian distribution. The fixation point distribution is concentrated in the photo center, but in the case of Curry and rice, it is biased towards the roux and in the case of Hamburger steak, it is bias in the direction of the sauce as shown in Figs. 9 and 10. These biases change the importance of local regions on the concentric circle. Therefore, we consider that the proposed method can select a feature extraction region with high influence on estimating the attractiveness.

V. Conclusion

This paper proposed a method for estimating the attractiveness of food photos. The proposed method extracts image features from the local regions selected based on the gaze information and estimates the attractiveness of a food photo by learning regression parameters. Also, we conducted a

preference experiment by subjects which compared pairs of food photos and measured their gaze. Through an evaluation experiment, we showed the effectiveness of extracting image features from outside the gaze regions rather than inside them, and confirmed that the effect of selecting the image feature extraction regions appears in the color feature rather than in the shape feature.

Future work includes increasing the number of food categories, improving the division of image and the analysis of gaze, introducing other kinds of image features, and selecting image feature extraction regions even when the view point is not given.

ACKNOWLEDGMENT

Parts of this research were supported by JSPS Grant-in-Aid for Scientific Research and MSR-Core12 program. We would like to thank the subjects for participating in the experiment.

REFERENCES

[1] M. Nishiyama, T. Okabe, I. Sato, and Y. Sato, "Aesthetic quality classification of photographs based on color harmony," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 33–40.

[2] X. Tian, Z. Dong, K. Yang, and T. Mei, "Query-dependent aesthetic model with deep learning for photo quality assessment." *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2035–2048, 2015.

[3] T. Kakimori, M. Okabe, K. Yanai, and R. Onai, "A system to support the amateurs to take a delicious-looking picture of foods," in *SIGGRAPH Asia 2015 Mobile Graphics and Interactive Applications*, 2015, p. 28:1.

[4] K. Takahashi, K. Doman, Y. Kawanishi, T. Hirayama, I. Ide, D. Deguchi, and H. Murase, "Estimation of the attractiveness of food photography focusing on main ingredients," in *Proceedings of the 9th Workshop on Multimedia for Cooking and Eating Activities in conjunction with the 26th International Joint Conference on Artificial Intelligence*, 2017, pp. 1–6.

[5] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proceedings of the 31st International Conference on Machine Learning*, 2014, pp. 647–655.

[6] S. Shimojo, C. Simion, E. Shimojo, and C. Scheier, "Gaze bias both reflects and influences preference," *Nature Neuroscience*, vol. 6, no. 12, pp. 1317–1322, 2003.

[7] Y. Sugano, Y. Ozaki, H. Kasai, K. Ogaki, and Y. Sato, "Image preference estimation with a data-driven approach: A comparative study between gaze and image features," *Journal of Eye Movement Research*, vol. 7, no. 3, pp. 5:1–5:9, 2014.

[8] R. Matsumoto, H. Yoshimura, M. Nishiyama, and Y. Iwai, "Feature extraction using gaze of participants for classifying gender of pedestrians in images," in *Proceedings of the 2017 IEEE International Conference on Image Processing*, 2017, pp. 3545–3549.

[9] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics*, vol. 23, no. 3, pp. 309–314, 2004.

[10] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[11] C. Michel, A. T. Woods, M. Neuhäuser, A. Landgraf, and C. Spence, "Rotating plates: Online study demonstrates the importance of orientation in the plating of food," *Food Quality and Preference*, vol. 44, pp. 194–202, 2015.

[12] L. L. Thurstone, "Psychophysical analysis," *American Journal of Psychology*, vol. 38, no. 3, pp. 368–389, 1927.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.