Analysis and Prediction of Attractive Fonts on Title-overlaid Food Images

Nanami TAKAGI
Nagoya University
Nagoya, Aichi, Japan
takagin@cs.is.i.nagoya-u.ac.jp

Haruya KYUTOKU
Aichi University of Technology
Gamagori, Aichi, Japan
kyutoku-haruya@aut.ac.jp

Keisuke DOMAN
Chukyo University
Toyota, Aichi, Japan
kdoman@sist.chukyo-u.ac.jp

Takahiro KOMAMIZU
Nagoya University
Nagoya, Aichi, Japan
taka-coma@acm.org

Ichiro IDE
Nagoya University
Nagoya, Aichi, Japan
ide@i.nagoya-u.ac.jp

Abstract

Title-overlaid images are useful as thumbnails on social media, where users prefer concise information to share and watch contents. Focusing on food contents, we aim to support creation of attractive title-overlaid food images to attract viewers' attentions. This paper first analyzes the effect of font styles of the title on the attractiveness of title-overlaid images via preference experiments, and creates a dataset. Next, we propose an given food image and the text feature of its food title to predict attractive fonts. This is the first attempt to build a model that selects an attractive font by considering the impression of the food image and its title.

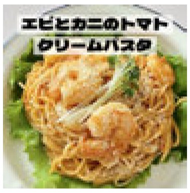
2 Attractiveness Analysis

We performed an analysis to clarify the relationship between the title font and the attractiveness of title-overlaid images. Details are described in this Section.

2.1 Creation of title-overlaid food images

Before collecting food images, we first selected 5 food categories: *rice bowl*, *omelet rice*, *pasta*, *salad*, and *soup*, considering the variation of foods in terms of appearance and composition of ingredients. In addition, a representative type of food was selected for each food category. Next, we collected, in total, 515 square

Figure 2: Fonts from Google Fonts[2] used in our experiment.



Noto Sans Japan Dela Gothic One Kaisei Decol

Figure 3: Examples of title-overlaid food images.

food images without overlaid titles by collecting 103 food photos for each food category on Instagram[1].

The collection was made in a way that there was as small bias as possible in the types of foods within each category. Also, the colors and positions of the plates and the shooting angles were also fixed as much as possible. We also checked if the structure and brightness of the images were appropriate and their food characteristics were clearly recognizable. We then selected food images with Japanese titles that included basic characters (*hiragana*, *katakana*, and *kanji*) and did not include symbols or numbers.

For the fonts for the overlaid titles, we chose 7 fonts shown in Fig. 2, considering the following attributes:

- Thickness: Degree of thickness of the strokes
- Roundness: Degree of roundness of curves and edges of the strokes
- Serif / Non-serif: Presence of extra strokes at the ends of strokes

Finally, we overlaid the food title on each food image with each font, as shown in Fig. 3. To clearly capture the relationship between the food images and fonts, overlay conditions were fixed so that the position and size of the food and the title should be uniform. As a result, we obtained 3,605 title-overlaid images (103 images × 5 food categories × 7 fonts).

2.2 Assignment of attractiveness scores

To assign attractiveness scores to the title-overlaid food images created in Sect. 2.1, we conducted a preference experiment. 30 panelists were asked to choose 4 out of 7 fonts based on their attractiveness for each food image; 2 as more and 2 as less attractive. We assigned $J = 3$ panelists for each image set of 7 fonts.

We used an ordinal scale score to represent the attractiveness of each font to the food image based on this preference experiment. Specifically, attractiveness score S_{if} of font f for food image i was calculated by averaging score S_{ifj} assigned to font f by panelist j ($1 \leq j \leq J$) for each food image i, as:

$$S_{if} = \frac{1}{J} \sum_{j=1}^{J} S_{ifj}, \qquad (1)$$

where the attractiveness score S_{ifj} is defined as:

$$S_{ifj} = \begin{cases} +5 & \text{if } (i, f) \text{ is judged as more attractive by } j, \\ 0 & \text{if } (i, f) \text{ is judged as less attractive by } j, \\ +1 & \text{otherwise.} \end{cases} \qquad (2)$$

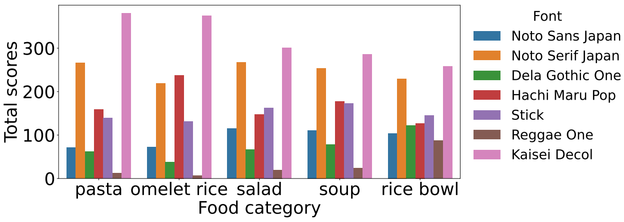


Figure 4: Experimental results: Total attractiveness scores for each font and food category.

2.3 Results and discussion

Results are shown in Figs. 4 and 5. Figure 4 shows the total attractiveness scores for all the images for each font and food category. Figure 5 shows some representative examples chosen as attractive.

Kaisei Decol was preferred for *pasta* and *omelet rice*. This could be because the elegant, graceful, and fashionable impressions of the font matched the fashionable impression of Western foods such as *pasta* and *omelet rice*. *Hachi Maru Pop* with a round impression, could have been preferred for *omelet rice* with a round shape. In the category of *rice bowls*, many of the foods gave the impression of being spicy, heavy, and with stamina, and thus the stinging *Reggae One* was most likely preferred. Furthermore, with respect to attractive fonts for each type of food, *Reggae One* was preferred for *chige soup*, which has a spicy impression. For *pasta*, *Noto Serif Japan* was judged as attractive for \Japanese style" foods, as was the case with *miso soup* and *Chinese soup*. We noticed round-shaped ingredients, such as shrimp and scallops, were judged as attractive for *Hachi Maru Pop*.

These results agree well with the design concepts of fonts in terms of the food-font relationship.

[1] `https://www.instagram.com/` [Accessed: April 1, 2025]
[2] `https://fonts.google.com/` [Accessed: April 1, 2025]

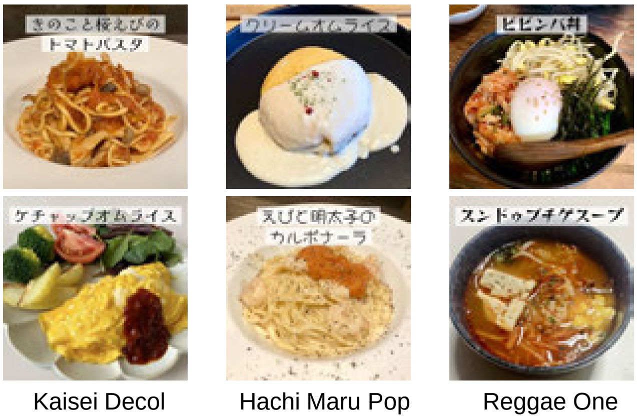


Figure 5: Experimental results: Examples of images chosen as attractive for different fonts.

Figure 6: Proposed font selection framework.

3 Proposed Model for Font Selection

We propose a multimodal model that calculates the attrativeness score by considering both image features of a food image and text features of a food title. An overview of our model is shown in Fig. 6.

For the food image, a pre-trained CNN-based model extracts image features as a D_i-dimensional vector. It is input to Fully Connected Layers (FCL) with Rectified Linear Unit (ReLU) activation, and converted to a D_c-dimensional vector. For the food title, BERT [6] extracts a D_t-dimensional vector as generic features output from an input [CLS] token. The vector is then converted to a D_c-dimensional feature through FCL+ReLU. Finally, both features are concatenated into a $2D_c$-dimensional vector, which is processed through FCL+ReLU, and converted into an N-dimensional vector.

This process, utilizing FCL+ReLU at multiple stages, effectively captures both the modality information of the food image and the food title, enhancing the model's capability to select appropriate fonts.

4 Evaluation Experiments

In this Section, we evaluate our framework for font selection described in Sect. 3.

4.1 Conditions

We built a font-attractiveness image dataset based on the results of the preference experiments in Sect. 2. It contains 515 original food images in total (103 images for 5 food categories). Each original image was associated with a 7-dimensional vector (ground-truth) of total attractiveness scores (Sect. 2.2), which represents how much each font was preferred within the 7 fonts (Fig. 2) in the preference experiment. The dataset was divided into 2 subsets in a 7:3 ratio by stratified sampling; one for training, and the other for testing.

We used Mean Squared Error (MSE; Range $[0, 25]$) and Accuracy (Acc) as evaluation metrics. Acc is the percentage of the original food images that the proposed method could correctly predict the ground-truth fonts. We evaluated Top-1 and Top-3 Acc.

We used 2 pre-trained CNN models: ResNet18 [7] and VGG16 [8], for image feature extraction, and a pretrained Japanese BERT model[2] [9] for text feature extraction. The dimensions of the feature vectors were $D_i = 512$ for ResNet18, $D_i = 4,096$ for VGG16, $D_c = 4,096$, $D_t = 768$, and $N = 7$. We trained each model using Adaptive Moment Estimation (Adam) [10] for optimizing the parameters, with a learning rate of $\alpha = 0.001$ and a batch size of 32. Data augmentation was applied to the training data in the order of, 1) random horizontal flip, 2) random cropping in the range of 70–100% of the image area, and 3) random rotation in the range of 0–60°.

4.2 Results of quantitative evaluation

Table 1 shows the quantitative evaluation results of the font selection task using our model. ResNet18 showed excellent performance on the training data, whereas that on the test data decreased significantly, which can be attributed to the high expressive ability of its features that make it prone to overfitting. VGG16 performed poorly in the training data compared to ResNet18, but showed a relatively small performance loss and slightly higher accuracy on the test data. This would be because the simpler VGG16 model represented only rough features and tended to be less prone to overfitting.

For reference, we also show the evaluation results with models trained with CNN and BERT individually in Table 1. Our model combining them achieved significantly higher performance compared to the individual models. Specifically, when trained with CNN alone, the model relied on partial information captured from the image, which led to limited accuracy. On the other hand, BERT alone excelled at capturing the meaning and context of the text, maintaining relatively high accuracy, although the accuracy was somewhat lower than our model. This shows that our model that combined VGG16 and BERT was the most effective,

[2] `cl-tohoku/bert-base-japanese-char-whole-word-masking`

Table 1: Quantitative evaluation results: MSE on Training / Testing data and Top-1/-3 Acc. † indicate compared methods.

Model	MSE # (Train / Test)	Acc [%] " (Top-1 / -3)
ResNet18+BERT	**0.075** / 1.540	52.7 / 85.3
VGG16+BERT	0.467 / **1.210**	**63.3** / **92.0**
ResNet18 [7] only†	0.100 / 1.733	49.0 / 84.0
VGG16 [8] only†	0.811 / 1.735	52.0 / 83.7
BERT [6] only†	0.761 / 1.356	55.0 / 88.0

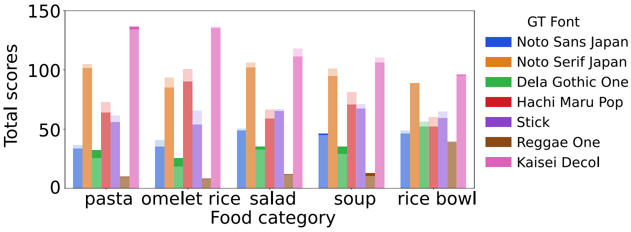


Figure 7: Distributions of the predicted results by the proposed method using VGG16.

maintaining high performance while suppressing overfitting, and providing the most stable results.

We also show the distribution of Ground-Truth vectors (GT) together with the predictions by our method (VGG16-based model, the best in Table 1) in Fig. 7. These distributions are very similar, indicating that our method was capable of adequately learning the relationship between food images and their titles, and consequently, the model accurately captured the distribution of font preference for each food.

4.3 Results of qualitative evaluation

Some examples of the font selection results by our method with VGG16 are shown in Fig. 8. As mentioned in Sect. 2.3, our model selected fonts based on the impression of foods, such as *Reggae One* for foods with spicy impression. It shows the ability to appropriately learn and predict the relationship between the impression of food images and titles and their font styles.

However, in cases such as food images for *Chinese cabbage cream soup* or *soy sauce butter corn potage*, the ground-truth font was *Kaisei Decol*, but the proposed model often predicted *Noto Serif Japan* as a more attractive choice. This discrepancy can be attributed to the fact that human evaluation tends to emphasize elements of Western cuisine, such as *cream soup* or *corn soup*, which give an impression of elegance and sophistication, leading to *Kaisei Decol* being rated as more attractive. In contrast, the proposed model strongly relied on keywords in the food titles, such as *Chinese cabbage* or *soy sauce*, which were more strongly associated with Japanese cuisine.

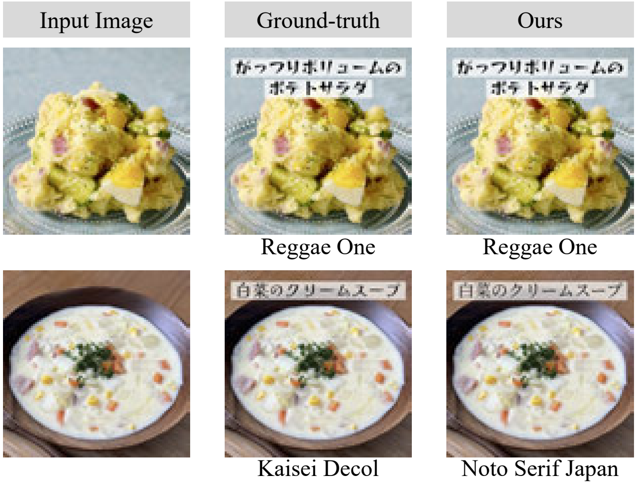


Figure 8: Examples of font selection results.

In other words, while human evaluations are based on the overall appearance and impression of the food image, our model might be biased toward specific words in the food title. To address this issue, it is necessary to design an architecture that effectively integrates both visual and textual information, enabling a balanced and comprehensive assessment.

5 Conclusion

We analyzed the attractiveness of title-overlaid food images posted on social media, and created a dataset annotated with font preferences by human panelists. Then, based on it, we made a multimodal font selection model. Quantitative and qualitative evaluations showed the effectiveness of our model.

Future work includes 1) increasing experimental data with more various overlaying conditions such as text positions, sizes and colors, 2) increasing panelists for assignment of attractiveness scores, 3) applying regularization techniques to avoid overfitting, 4) introducing an architecture that properly captures the key features in both image and text in terms of attractiveness, and 5) more comprehensive evaluations by comparing the proposed method with state-of-the-art multimodal LLMs.

Acknowledgments

This work was partially supported by JSPS KAKENHI #22H00548.

References

[1] J. Kang, D. Haraguchi, S. Matsuda, A. Kimura, and S. Uchida, "Shared latent space of font shapes and

their noisy impressions," in *Proc. 28th Int. Conf. Multimed. Model.*, June 2022, vol. 2, pp. 146–157.

[2] K. Lei, Y. Fei, W. Kai, A. S. Mohamed, G. Lluis, F. Alicia, V. Ernest, and K. Dimosthenis, "GRIF-DM: Generation of rich impression fonts using diffusion models," in *Proc. 27th Eur. Conf. Artif. Intell.*, Aug. 2024, pp. 226–233.

[3] S. Matsuda, A. Kimura, and S. Uchida, "Font generation with missing impression labels," in *Proc. 26th Int. Conf. Pattern Recognit.*, Aug. 2022, pp. 1400–1406.

[4] Y. Kasai and H. Sato, "Study of designing technique for sizzle feeling of font (in Japanese)," in *Proc. 64th Annu. Conf. Jpn. Soc. Study Design*, June 2017, pp. 186–187.

[5] X. Chen and W. Tang, "Study on the influence of font type on the aesthetics evaluation of food packaging," in *Proc 15th Int. Conf. Human System Interact.*, July 2022, 5 pages.

[6] D. Jacob, C. Ming-Wei, L. Kenton, and T. Kristina, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. N. Am. Chapt. Assoc. Comput. Linguist.*, May 2019, pp. 4171–4186.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. 2016 IEEE Conf. Comput. Vis. Pattern Recognit.*, Dec. 2016, pp. 770–778.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent.*, May 2015, 14 pages.

[9] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, "Pre-training with whole word masking for Chinese BERT," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, no. 11, pp. 3504–3514, Nov. 2021.

[10] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Comput. Res. Reposit. arXiv Preprint*, , no. arXiv:1412.6980, Dec. 2015.