Applied Human Factors and Ergonomics (AHFE2025), Vol. 199, 2025, 1263–1269
https://doi.org/10.54941/ahfe1006937

Estimating Product Attributes Relevant to Purchase Decisions From Images in C2C Marketplaces

Kohei Otake[1] and Yoshihisa Shinozawa[2]

[1]Faculty of Economics, Sophia University, 7–1 Kioi-cho, Chiyoda-ku, Tokyo 102-8554, Japan
[2]Faculty of Science and Technology, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama-shi, Kanagawa 223-8522, Japan

ABSTRACT

In recent years, consumer-to-consumer (C2C) online flea markets, which are platforms where individuals buy and sell goods directly, have grown rapidly. Prior studies suggest that consumer behavior on C2C platforms differs from that on business-to-consumer platforms, prompting research that leverages multimodal information, such as images and text. Among these modalities, image analysis plays a key role in revealing visual cues that influence purchase decisions. Manually annotated labels are often used to ensure interpretability; however, large-scale annotation is costly and labor intensive, limiting scalability. This study addresses this issue by developing deep-learning models that automatically estimate the product attributes that affect purchase decisions. We analyzed the product images of tops from a fast-fashion brand posted on a company-operated C2C platform. Using thumbnail images, we built models to predict five visual attributes: (1) Packaged, (2) Folded, (3) Characters, (4) Official Website Image, and (5) Size. Four architectures, namely ResNet, EfficientNet, ConvNeXt, and Swin Transformer, were compared in terms of accuracy. All classification tasks achieved an accuracy of over 90%, with the best-performing model varying by attribute. These results demonstrate that deep-learning-based automatic annotation can effectively reduce labeling costs and support scalable consumer behavior research on C2C platforms.

Keywords: Consumer-to-consumer platform, Deep learning, Automatic annotation

INTRODUCTION

Consumer-to-consumer (C2C) e-commerce (EC) platforms have attracted increasing attention in recent years, driven by advancements in information and communication technology, increasing prices, and growing awareness of sustainability. According to the FY2024 Survey on Electronic Commerce (Ministry of Economy, Trade and Industry, 2025), the domestic C2C-EC market in Japan reached JPY 2.5269 trillion in 2024, marking a 1.82% increase from the previous year. This figure indicates that the C2C-EC market has grown to a scale comparable to that of the digital business-to-consumer (B2C)-EC sectors. The same report also notes that in China, which holds the world's largest B2C-EC market, the reuse (secondhand) EC market

reached 548.7 billion RMB in 2023—a year-on-year increase of 14.3%—and, although growth is moderating, it is expected to remain steady in the coming years.

C2C-EC platforms are often referred to as "flea market apps." Similar to offline flea markets, sellers list items on these platforms by uploading product images, descriptions, and prices. Buyers select their preferred items and complete payments through the platform. The seller ships the item to the buyer and the transaction is completed when the buyer receives the product. Among these elements, product images, particularly thumbnail images, which serve as the first visual cues encountered by consumers, are known to have a significant impact on purchase decisions (Wang et al., 2018).

Previous studies focusing on thumbnail images include those by Sato and Tamura (2019) and Zhang et al. (2021), which analyzed viewer preferences on video-sharing platforms using thumbnail images, as well as Miyamoto (2022), who applied machine learning to examine the relationship between thumbnail characteristics and access frequency. In addition, Kozu et al. (2024) experimentally evaluated the sensitivity differences of thumbnail images on online tourism video platforms. These studies have consistently demonstrated that thumbnail images are the initial determinants of user selection processes and play an important role in product comparisons and evaluations.

In our previous study (Iwanade et al., 2025), we sought to identify the factors influencing purchasing behavior in online flea markets by constructing a binary classification model to distinguish between sold and unsold items. The model utilized both metadata, such as price and shipping methods, and manually annotated labels that were applied to product thumbnail images as features. The results indicated that items priced between JPY 1,000 and 2,200 and those listed as "new" or "unused" had higher probabilities of being sold. In terms of image attributes, items for which official website images or images of packaged products were used as thumbnails were more likely to be purchased.

However, studies in this field often rely on manual annotations to ensure interpretability, which poses significant scalability limitations. Conducting large-scale manual labeling of numerous images is both costly and time consuming, making it difficult to construct sufficiently large training datasets. In our previous study, the dataset consisted of only 3,000 images, suggesting that further analysis of consumer behavior and decision-making processes requires more efficient annotation methods.

Therefore, in the present study, we aimed to automate the annotation of attributes that influence purchase decisions using deep-learning models. Specifically, we constructed models to predict five attributes: (1) whether the product is packaged, (2) whether the product is folded, (3) whether the product features a character or franchise, (4) whether an official website image is used, and (5) whether the image contains text indicating the clothing size. We used the product images of tops from a fast-fashion brand posted by users on a company-operated C2C platform as input thumbnails. We developed four deep-learning models for each attribute—ResNet (He et al., 2016), EfficientNet (Tan and Le, 2019), ConvNeXt (Liu et al., 2022),

and Swin Transformer (Liu et al., 2021)—and evaluated their predictive performances through an accuracy comparison.

DEVELOPMENT OF DEEP-LEARNING MODELS FOR AUTOMATIC ESTIMATION OF PURCHASE-RELEVENT PRODUCT ATTRIBUTES

This section describes the development of the deep-learning models designed to predict the five attributes that were identified as influencing purchase decisions in previous research.

Dataset

We used the Mercari Dataset provided by the Informatics Research Data Repository of the National Institute of Informatics, which is related to the online flea market application Mercari (Mercari, Inc., 2023).

The dataset used in this study is an extended version of that used in our previous study (Iwanade et al., 2025). Specifically, whereas the previous study utilized an annotated dataset comprising 3,000 records, we constructed an expanded dataset containing 10,000 manually annotated records for the present analysis. Two graduate students who were part-time research assistants performed the annotation task. Both annotators were experienced users of Mercari and were deemed to possess sufficient knowledge to label product images based on the labeling framework established in our previous work.

As the target of the analysis, we focused on the fashion category, selecting men's and women's tops, which had the highest number of transactions. Among the listed items, we targeted the two brands with the largest transaction volumes, namely a fast-fashion brand and its sister brand, and limited the price range to between JPY 300 and JPY 10,000 to account for outliers. As in the previous study, the dataset contained 5,000 items labeled as "sold" and 5,000 labeled as "on sale," representing two transaction states at a given point in time.

In addition to product data, such as item descriptions and shipping information, the dataset included thumbnail images associated with each listing. In this study, we specifically focused on the thumbnail images and five attributes identified in the previous research as key factors influencing purchase occurrence. Table 1 summarizes the definitions of these attributes and their respective frequencies across all images.

Table 1: Overview and frequency of labels.

Attribute Name	Criterion Description	Number of Labels
Packaged	The clothing item is photographed in a packaged state.	563
Folded	The clothing item is photographed in a folded state.	482

Continued

Table 1: Continued

Attribute Name	Criterion Description	Number of Labels
Characters	The clothing item features a collaboration with a character or franchise.	987
Official Website Image	The image is obtained directly from an official website.	1,243
Size	The image includes visible text indicating the clothing size.	850

Model Development

To automate the annotation of the visual attributes described in the previous section, four deep-learning models specialized for image-recognition tasks were implemented in this study: ResNet, EfficientNet, ConvNeXt, and Swin Transformer. ResNet, EfficientNet, and ConvNeXt are convolutional neural network (CNN)-based architectures, whereas the Swin Transformer is based on the Transformer architecture.

We constructed binary classification models for each of the five labels identified as factors influencing purchase occurrence in the previous research and evaluated their predictive performance. To determine whether an item in an image is folded, it is necessary to detect the item and then determine whether it is in a folded state. The images targeted in this study contained one or two items. Furthermore, most items were positioned near the center of the image. Therefore, we omitted the object-detection process and instead treated it as a classification problem in which the state of an item was predicted as a label. For model training, images containing the target label were extracted and negative samples (images without labels) were undersampled to balance the dataset. The resulting data were then divided into training (80%) and test (20%) sets, maintaining a 1:1 ratio of positive to negative samples.

Based on the predictions obtained for the test data, we computed the following evaluation metrics to compare the model performances:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

$$F1\text{-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}, \tag{2}$$

where TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively (Equation (1, 2)).

All models utilized pretrained models from the timm package of PyTorch (Hugging Face), ResNet (*ResNet50*), EfficientNet (*tf_efficientnetv2_b0*), ConvNeXt (*convnext_tiny.fb_in22k*), and Swin Transformer (*swin_tiny_patch4_window7_224*). These pretrained models have been trained on ImageNet (ImageNet). The models were fine-tuned using the dataset developed in this study. The training parameters were

set as follows: input image size = 224 × 224, batch size = 32, maximum epochs = 200, early stopping patience = 10, and dropout rate = 0.5.

Data augmentation was applied to the training images, as summarized in Table 2.

Table 2: Overview of data augmentation for training images.

Processing Name	Description
Resize	Resize both height and width to 224 pixels
RandomHorizontalFlip	Horizontally flip the image with a 50% probability
RandomRotation	Randomly rotate the image within ±10 degrees
RandomResizedCrop	Randomly crop 80%–100% of the image and resize it to 224 × 224 pixels
ColorJitter	Randomly adjust brightness, contrast, and saturation within a range of 0.8–1.2×
ToTensor	Convert the image into a tensor and scale pixel values to [0, 1]

MODEL EVALUATION AND DISCUSSION

This section presents the predictive performances of the models developed to estimate the five attributes that influence purchase occurrence. In all models, early stopping was triggered within 40 epochs. Notably, for three architectures, namely EfficientNet, ConvNeXt, and the Swin Transformer, the final models were obtained within 10 epochs, except for the EfficientNet model when predicting the Official Website Image attribute, which converged at 11 epochs. In contrast, the ResNet model required up to 28 epochs, indicating relatively slow convergence compared with the other three architectures.

Table 3 summarizes the final model accuracies for each task and Table 4 presents the corresponding F1-scores. The highest performance values for each attribute are highlighted. The reported F1-scores were re-evaluated using the best-performing models saved after early stopping.

Table 3: Final model accuracies for each attribute.

Architectures	Packaged	Folded	Characters	Official Website Image	Size
ResNet	0.9013	0.8705	0.9859	0.9867	0.9176
EfficientNet	0.9266	0.8705	0.9920	0.9735	0.9353
ConvNeXt	0.9241	0.8860	0.9819	0.9912	0.9382
Swin Transformer	0.9342	0.8860	0.9819	0.9912	0.9265

Table 4: Final F1−scores for each attribute.

Architectures	Packaged	Folded	Characters	Official Website Image	Size
ResNet	0.9277	0.8367	0.9899	0.9868	0.8982
EfficientNet	0.9323	0.8657	0.9939	0.9778	0.9244
ConvNeXt	0.9453	0.9043	0.9857	0.9778	0.9348
Swin Transformer	0.9463	0.8973	0.9899	0.9912	0.9459

As shown in Tables 3 and 4, the overall performance of the developed models indicates that all five target attributes were predicted with high accuracy. Among these, the label *Official Website Image*, which indicates whether the image originated from an official source, achieved exceptionally high predictive performance across all models, with F1-scores exceeding 97%. In contrast, the label *Folded*, indicating whether the clothing item was photographed in a folded state, exhibited relatively lower predictive accuracy compared with the other attributes. The models that achieved the highest F1-scores for each task are highlighted in Table 4. Specifically, the Swin Transformer exhibited the best performance for *Packaged*, *Official Website Image*, and *Size*, whereas ConvNeXt achieved the highest accuracy for *Folded*, and EfficientNet performed the best for *Characters*. Notably, for the *Folded* attribute, there was a performance gap of approximately 7% between the lowest-performing model (ResNet) and highest-performing model (ConvNeXt), which was larger than the differences observed in other tasks. Although the variations among the other tasks were smaller, this result suggests that certain architectures may have strengths and weaknesses depending on the inherent visual characteristics of each attribute. These findings indicate that the differences in feature extraction mechanisms between CNN- and Transformer-based architectures may contribute to their varying performance across visual attribute types.

CONCLUSION

This study aimed to estimate the visual attributes that influence purchase decisions in online flea markets automatically using deep-learning models. Specifically, we focused on the tops of a fast-fashion brand listed by users and constructed models to predict five attributes from thumbnail images: (1) Packaged, (2) Folded, (3) Characters, (4) Official Website Image, and (5) Size.

Four deep-learning architectures, namely ResNet, EfficientNet, ConvNeXt, and Swin Transformer, were trained and compared for each label. The results indicated that all five attributes could be predicted with high accuracy. In particular, the Official Website Image attribute achieved an F1-score of over 97% across all models, while the Folded attribute exhibited slightly lower accuracy, with a performance gap of approximately 7% between architectures. These results suggest that each model has distinct strengths and weaknesses in terms of feature extraction.

Overall, the findings indicate that deep-learning-based automatic annotation can effectively reduce the cost of manual labeling and enable

large-scale behavioral analysis on C2C platforms. Future work should explore multilabel prediction and multimodal approaches that integrate visual and textual data to clarify consumer decision-making processes in C2C marketplaces further.

ACKNOWLEDGMENT

We used the Mercari Dataset provided by Mercari, Inc. via the IDR Dataset Service of the National Institute of Informatics. This work was supported by JSPS KAKENHI (grant number 24K05159).

REFERENCES

Hiroki Kozu, Kosuke Saito, Shunsuke Nakamura, Riho Kondo and Yasuhiro Tsujimura. (2024), "A study on sensibility evaluation of thumbnails on online video sharing services for tourism targeting Japanese elderly people," Journal of Global Tourism Research, 9(1), 13–21.

Hugging Face, https://huggingface.co/timm.

ImageNet, https://www.image-net.org/.

Iwanade, E., Shinozawa, Y., & Otake, K. (2025). "Identifying Purchasing Factors in Online Flea Markets Considering Thumbnail Images," Journal of Data Science and Intelligent Systems, Online First, 9pages.

K. He, X. Zhang, S. Ren and J. Sun. (2016), "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770–778.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 10012–10022).

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). "A ConvNet for the 2020s," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 11976–11986).

Mercari, Inc. (2023): Mercari Dataset. Informatics Research Data Repository, National Institute of Informatics. (dataset). https://doi.org/10.32130/idr.17.1

Ministry of Economy, Trade and Industry (2025), "FY2024 Electronic Commerce Market Survey Report." (in Japanese).

Ryosuke Sato and Ryoichi Tamura. (2019). "Study on Thumbnail Images and Titles Selected by Viewers in TouTuber's Video," International Journal of Affective Engineering, 18(1), 139–145.

S. Zhang, T. Aktas and J. Luo. (2021). "Mi YouTube es Su YouTube? Analyzing the Cultures using YouTube Thumbnails of Popular Videos," 2021 IEEE International Conference on Big Data, 4999–5006.

Tan, M., & Le, Q. V. (2019). "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," In Proceedings of the 36th International Conference on Machine Learning (ICML), PMLR 97:6105–6114.

Wang, Y., Guo, Y., & Song, J. (2018) "Using image-based and text-based information for sales prediction: A deep neural network model," Americas Conference on Information Systems 2018: Digital Disruption, AMCIS 2018).

Yukinobu Miyamoto. (2022). "An Approach to Classify Thumbnail Images on Video Sites by the Number of Accesses," 37th International Technical Conference on Circuits/Systems, 194–197.