Visual Analytics for Brand Identity in Packaging Design

Asaya Shimojo

Abstract

Many brands traditionally rely on qualitative methods to design their product packaging, leaving uncertainties about the consistency of brand identity across different packages. This study leverages machine learning to quantitatively extract and analyze design elements that resonate with consumers' perception of brand identity. Specifically, we employed Grad-CAM, an interpretative method for Convolutional Neural Networks (CNNs), to identify crucial visual features—termed Visual Identities—within the middle layers of a model trained on specific brand package images. These features were analyzed to determine their influence on package classification and their alignment with human perception of brand identity. Our findings demonstrate that the machine learning approach approximates human perception closely, providing a novel quantitative method to enhance and maintain brand identity. Additionally, we quantified the contribution of each identified Visual Identity to overall brand recognition, offering a more systematic approach to understanding and preserving a brand's distinctiveness that has traditionally been handled qualitatively.

Introduction

Brand identity (BI) serves as the embodiment of a company's values and characteristics, with visual elements such as logos, colors, and design styles playing crucial roles in its expression [1], [2]. Traditionally, package design evaluation has relied heavily on qualitative methods, making it challenging to maintain BI consistency across different products and objectively assess design elements' impact on brand recognition. This study developed a novel quantitative approach to analyzing BI in product packaging using machine learning.

The previous research has shown that package design significantly influences consumers' emotions and perceptions through its BI, including colors, shapes, materials, logos, and text [2]. The visual attractiveness of packaging has been

KONICA MINOLTA, Inc., Tokyo, Japan

demonstrated to draw consumer attention and form favorable product impressions [3]. BI consistency has been proven to improve purchase intent and brand loyalty [4], though maintaining this consistency while adapting to changing trends presents a significant challenge [5]. Notably, interviews with art directors and other professionals revealed they have a narrow range of acceptance regarding BI consistency when making visual identity changes. Some consumers with high aesthetic sensitivity also demonstrated similar concerns about visual identity modifications [5]. However, these professional assessments are often based on subjective preferences and emotions, highlighting the need for more objective, quantitative methods for evaluating visual identity elements.

The aim of this study is to develop a machine learning-based framework that can quantify the specific visual elements within packaging design that contribute to brand recognition. Specifically, we employed Grad-CAM visualization techniques to extract the key visual features that the model relied upon to distinguish between the different brand identities. By quantifying the contribution of these individual design elements, this research offers a data-driven approach to managing BI through packaging. Previous studies have examined BI by focusing primarily on logos [6], [7], but our approach extends beyond logos to comprehensively analyze all visual elements that contribute to brand recognition, including object edges, color contrasts, and other design features [8].

A convolutional neural network (CNN) is structured to hierarchically extract low-level features such as edges, colors, and shapes from images, and progressively transform them into higher-level semantic representations. Consequently, differences in visual elements such as logos, curves, and surface glossiness within packaging are abstracted in a layer-wise manner inside the model and are learned as feature representations that contribute to brand classification. The novelty of this study lies in leveraging this hierarchical structure of CNNs to extract visual contribution elements from deep-layer activation maps.

Furthermore, in addition to evaluating model performance using standard accuracy metrics, we conducted a comparative experiment with eye-tracking data to assess the alignment between the model's output and human visual perception. While the visualization of CNN decision-making processes has been established through techniques like Grad-CAM [9], [10], there remains a gap in validating whether these machine-identified features align with human visual attention patterns when recognizing brands [11]. Our study addresses this gap by directly comparing human eye-tracking data with CNN visualization results.

To ensure reproducibility and scalability, we also explicitly documented the dataset construction and preprocessing procedures, designing the framework to enhance its applicability in future research and practical implementations. Our approach enables not only the identification of key visual elements that contribute to brand recognition but also the quantification of each element's specific contribution to the overall brand identity, providing brand managers with concrete metrics to guide packaging design decisions.

Methodology

Model Construction

We developed a multi-class classification model to identify package designs from different brands. They used VGG16 pre-trained on ImageNet and a dataset comprising approximately 6,000 images from five writing instrument brands, categorized following Kapferer's luxury brand scale [12]: two luxury brands (Brand A and B), two general consumer brands (C and D), and one lesser-known brand (E).

Images were collected from the official websites and e-commerce platforms of each brand, with approximately an equal number of package images obtained per brand, resulting in a dataset of around 1,200 images in total. All images were resized to 224×224 pixels and normalized across RGB channels. During training, data augmentation techniques such as random cropping and horizontal flipping were applied to mitigate overfitting. The model was trained with 70% of the dataset images, and the remaining 30% was used for validation. The training and validation processes were conducted 2,000 times, shuffling the data each time, with a batch size of 32 and the number of epochs set to 10.

Model performance was evaluated by varying the neuron counts in the model's dense layer from 51 (10% of full capacity) to 512 (100%). Model optimization revealed that 512 neurons achieved 91.18% validation accuracy, while models with 384 neurons also performed well, achieving an accuracy of 87.47%. However, models with fewer neurons—256, 128, and 51—all showed validation accuracies below 80%.

Furthermore, when the training dataset was reduced, the model with 512 neurons showed varying accuracies: 62% with 600 images (10% of full data), 77% with 1,500 images (25%), 82% with 3,000 images (50%), and 89% with 4,500 images (75%). A training dataset of at least 4,500 images was considered necessary for optimal performance. This study used the largest model size and number of samples available (100% for both) because accuracy was given priority.

We then employed Grad-CAM [9], [10] to visualize the crucial regions in the input images that the CNN model focused on when making classification decisions. These visualizations are shown in Figure 1, which displays the areas that contributed to an image being classified as Brand A.

Figure 1. Brand A classification contribution area acquired by Grad-CAM.

Grad-CAM generates heatmaps by calculating weights based on the gradients of a specific convolutional layer output A^k with respect to the classification score y^c, and produces the heatmap as a weighted sum. Specifically, it is defined as:

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \partial y^c / \partial A_{ij}^k, \quad L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right)$$

where Z is a scalar for spatial averaging, and ReLU means blocking negative activations. This highlights the spatial regions that provide the basis for classification.

In applying Grad-CAM, we set the visualization target layer to block5_conv3, which is the final convolutional layer. This layer is highly abstract and responds to the entire object, making it suitable for evaluating the visual consistency of brands. Specifically, we calculated the gradients of the loss function for each class with respect to the output of this layer and generated heatmaps through a weighted sum passed through the ReLU function. In our ablation study, when this layer was disabled, the validation accuracy decreased by over 70%, indicating the layer's critical role in accurate brand recognition. This confirmed the importance of visualizing this layer to identify visual elements that affect brand recognition.

Model Evaluation

The performance of the proposed model was evaluated using validation accuracy, precision, recall, and F1 score. In the five-class classification task, the average F1 score was 0.89. Luxury brands (A and B) achieved an average accuracy of 0.92, general consumer brands (C and D) achieved 0.87, and the minor brand (E) yielded a relatively lower score of 0.83. An analysis using a confusion matrix revealed a notable trend of misclassification between Brand E and Brand C.

For training, we employed the Adam optimizer with a learning rate of 1e-4, a batch size of 32, and a total of 30 epochs. Early stopping was applied, terminating the training when the validation loss did not improve for three consecutive epochs.

In addition to VGG16, we conducted comparative validation with ResNet50 and Vision Transformer (ViT Base) under identical conditions. The same preprocessing and training settings (image size 224×224, Adam optimizer with learning rate 1e-4, batch size 32, Early Stopping) were applied to each model. Evaluation metrics included Validation Accuracy, Normalized Scanpath Saliency (NSS) score, and visualization results from Grad-CAM or Attention Maps.

While ResNet50 achieved relatively high accuracy (89.3%), its Grad-CAM heatmaps tended to spread over background regions, slightly reducing the visual interpretability of classification rationales. On the other hand, ViT demonstrated high visual concentration through its Attention Maps and showed good agreement with human fixation distributions, but the model size and training load resulted in higher implementation costs. Based on these results, we selected VGG16 for this study due to its balanced performance and high explainability.

Comparative Analysis of Human and Machine Perception

To investigate the alignment between model-extracted visual features and human recognition of BI, we conducted a comparative analysis using eye-tracking data. Six participants underwent standardized eye-tracking tests while viewing package images from the dataset.

The experimental conditions specified that each image was displayed for 5 seconds with free gaze, and blank screens were inserted before and after to randomize the initial fixation position. Participants ranged in age from 20s to 40s, and prior knowledge of the brands was not restricted. The participants' heads were fixed by a chinrest placed approximately 60 cm from the monitor screen. Images were displayed across the entire monitor screen (1920 × 1080 px resolution), and the experiment room brightness was kept constant at 500 lux.

Eye movements were recorded using a webcam (ELECOM UCAM-C750FBBK; resolution 1920 × 1080 px, frame rate 30 FPS, angle of view 66 degrees, 1/4-inch CMOS sensor) and GazeRecorder [13], a line-of-sight measurement software. After gazing at each image, participants responded with a confidence level ranging from 0% (definitely not a Brand A package) to 100% (definitely a Brand A package) regarding their certainty about the brand identity.

For the analysis, we used only human eye-tracking data with an average confidence level of at least 70% to ensure we were analyzing only package images that participants were confident were from Brand A. To compare the two different heat maps (one from human eye-tracking and one from Grad-CAM), we employed the

Jaccard Index in addition to the Normalized Scanpath Saliency (NSS) score (average: 1.35, SD: 0.45) to complementarily verify the spatial correspondence between CNN and human visual perception.

The comparison between the participants' gaze patterns and the Grad-CAM heatmaps yielded a Jaccard Index of 0.32 (SD = 0.12; Fig. 2), indicating a moderate correlation. This suggested that the CNN model's extraction of visual features aligns to some degree with how humans visually recognize BI, providing validation for the machine learning approach.

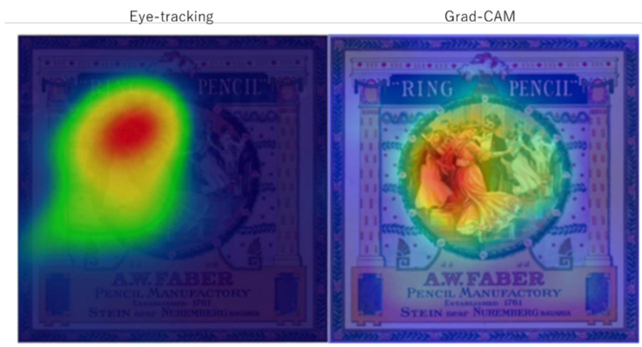

Figure 2. An example of a comparison of the features of the middle layer of the visualized CNN and the gaze trajectory.

However, it's important to note that Grad-CAM is based on "classification basis location" rather than visual attention, so there may be discrepancies with exploratory gaze or unconscious interest. Human fixation on textual information such as brand names may sometimes be overlooked by the model. To address this limitation, combining optical character recognition (OCR) with attention models could be effective.

Quantifying the Contribution of Visual Design Elements

Building on previous studies examining brand identity through logos [6], [7], we extended the analysis to comprehensive package design elements. By examining the activation patterns in the CNN's "block5_conv3" layer, we identified 40 significant filters out of 512, each capturing a distinct visual feature contributing to brand recognition.

For feature extraction, we analyzed the activation patterns of the output filters from the block5_conv3 layer, which consists of 512 channels. We calculated the correlation between the average activation value of each filter and the classification results, and selected 40 filters that showed the highest contribution to classification.

This enabled a quantitative analysis of visual elements contributing to brand identification from a perspective distinct from Grad-CAM.

To better understand each filter's functional importance, we conducted an ablation study where the weights of each filter in the middle tier were individually set to zero, and the impact of these changes on the model's overall testing accuracy was systematically evaluated. By disabling each filter, we quantitatively analyzed the extent to which the filter contributed to the model's ability to make decisions.

The analysis showed that the filters' contribution to Brand A recognition ranged from a maximum of 25% to a minimum of 0.12%. The key features included sharp object tips (13%; Fig. 3), product curve strength (11.2%), metal body reflectance (8.9%), and background-product contrast (8.1%). Of note was the 110th filter, which, when disabled, resulted in a 13% reduction in the ability to identify Brand A, indicating its crucial role in extracting the visual features characteristic of Brand A.

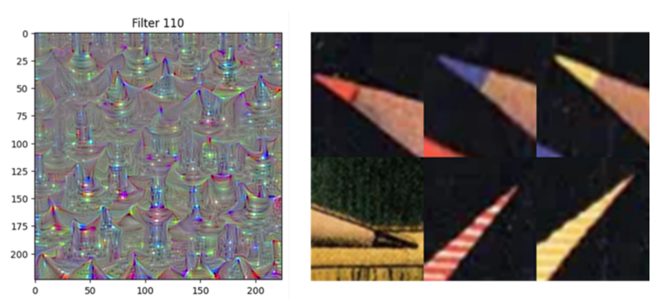

Figure 3. Activation Map of the 110th Filter Highlighting Response to Sharp Object Tips.

These findings offered a more nuanced understanding of the specific visual elements that drive brand recognition, going beyond the more holistic measures examined in previous research. Such insights are valuable in clarifying the key visual elements in brand identification and understanding how they affect the performance of the model. This process demonstrates that the extraction of visual identities from packaging and the calculation of their impact on brand identity can be effectively automated.

Discussion

The findings from this research demonstrate the potential of machine learning techniques, such as CNN and Grad-CAM, to quantify the impact of packaging design on BI. By identifying and validating the visual features that contribute to brand recognition, this study offers a more systematic and data-driven approach to managing BI through packaging.

The moderate correlation between human eye-tracking data and the CNN model's attention points suggests that the machine learning approach can approximate human perception to a reasonable degree. However, the discrepancies observed also indicate that human visual processing involves additional factors beyond the saliency-based features captured by the current CNN architecture.

While Grad-CAM effectively highlights crucial areas within an image, it tends to produce more diffuse heatmaps when background noise or multiple objects are present in the image. In package images particularly, where brand logos and product structures often overlap closely, clearly defining boundaries becomes challenging. To address these limitations, implementing weight-independent visualization methods like Score-CAM or introducing transformer-based attention analysis could be beneficial.

In our preliminary experiments, we compared VGG16 with deeper and lighter networks like ResNet50 and EfficientNet. ResNet50 achieved an accuracy of 89.3%, but its Grad-CAM heatmaps were somewhat diffuse, showing reduced locality in the basis for decisions. In contrast, Vision Transformer (ViT) models with attention visualization capabilities demonstrated better alignment with human fixation points in some cases, suggesting that model selection should consider compatibility with visualization methods.

The ability to extract and quantify the contribution of individual design elements provides brand managers with concrete metrics to guide packaging design decisions. This could enable more strategic adjustments to maintain brand consistency while adapting to evolving market trends and consumer preferences - a key challenge identified in prior literature [5].

This methodology builds upon Underwood and Klein's assertion [14] that visual identity plays a central role in BI communication, providing quantitative evidence for this relationship. The findings support earlier research showing that art directors and professionals have a narrow range of acceptance for visual identity consistency [5], while offering objective metrics for evaluating these changes.

However, the eye-tracking study was conducted with only six participants, which limits the statistical generalizability of the findings. Additionally, potential biases due to participants' cultural background or individual differences in visual cognition were not controlled. Future studies should include a more diverse and larger participant pool ($n \geq 30$) and consider screening procedures to account for variations in visual perception.

Future research should explore more advanced techniques, such as attention mechanisms and multi-dimensional scaling, to better align the machine's understanding of BI with human cognition. These methodologies could avoid the saliency bias, helping identify subtle yet crucial patterns that might be missed by traditional saliency-based approaches. They not only perform well but also align more closely with human cognitive processes, potentially making machine learning tools more intuitive and trustworthy for users in real-world applications. The research's practical implications extend to brand management, offering a systematic approach to extracting and quantifying VI elements. This aligns with previous findings on the importance of visual attractiveness in forming favorable brand impressions [3] while providing concrete metrics for maintaining consistency across product lines. The application of this methodology could aid brand managers in strategically adjusting packaging designs to cater to evolving consumer preferences and market trends, without compromising the core BI.

As technology continues to advance, this research represents a promising step towards automating the process of BI management, potentially revolutionizing how brands maintain visual consistency and effectiveness across their product portfolios. The outcomes of this study can be implemented as practical tools, such as a UX tool providing Grad-CAM visualization on a web interface to support designers' decision-making when modifying package compositions. In the future, it may even be possible to develop "individually optimized BI design" that connects user attributes with visual responses.

References

[1] Jean-Noël. Kapferer, "Strategic Brand Management." London:Kogan Page. 1992.

[2] D. A. Aaker, "Building Strong Brand." New York: The Free Press—A division of Simon & Schuster Inc. 1996.

[3] M. E. H. Creusen, and J. P. L. Schoormans, "The different roles of product appearance in consumer choice." Journal of Product Innovation Management, vol. 22, no. 1, pp. 63–81, 2005.

[4] M. Lindstrom, "Broad sensory branding." Journal of Product & Brand Management, vol. 14, no. 2, pp. 84–87, 2005.

[5] B. J. Phillips, E. F. McQuarrie, and W. G. Griffin, "How visual brand identity shapes consumer response." Psychology & Marketing, vol. 31, no. 3, pp. 225–236, 2014.

[6] R. Beise-Zee, "Brand equity retention after rebranding: a resource-based perspective." Journal of Brand Management, vol. 29, no. 2, pp. 208–224, 2022.

[7] A. Jakšić-Stojanović, and N. Šerić, "Brand Identity of Montenegro through Verbal and Visual Elements of its Logo." Journal of Marketing Development and Competitiveness, vol. 12, no. 4, pp. 134–144, 2018.

[8] V. U. Vinitha, D. S. Kumar, and K. Purani, "Biomorphic visual identity of a brand and its effects: a holistic perspective." Journal of Brand Management, vol. 28, no. 3, 272–290, 2021.

[9] R. R. Selvaraju, M. Cogswell, A. Das, Vedantam, R., Parikh, D., and Batra, D. "Grad-CAM: Visual explanations from deep networks via gradient-based localization." International Journal of Computer Vision, vol. 128, pp. 336–359, 2017.

[10] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2921–2929, 2016.

[11] J. An, and I. Joe, "Attention map-guided visual explanations for deep neural networks." Applied Sciences, vol. 12, no. 8, 3846, 2022.

[12] Jean-Noël. Kapferer, "Why are we seduced by luxury brands?." Journal of Brand Management, vol. 6, no. 1, pp. 44–49, 1998.

[13] GazeRecorder. Online Eye Tracking Software. [Online]. Available from: https://gazerecorder.com/ 2025.05.16

[14] R. L. Underwood, and N. M. Klein, "Packaging as brand communication: Effects of product pictures on consumer responses to the package and brand." Journal of Marketing Theory and Practice, vol. 10, no. 4, pp. 58–68, 2002. Series in Statistics, Springer-Verlag New York, 2005.