Impression Estimation Model for Suit Styles Using CNN Features *

Kimiaki Shinkai[1], Iori Tani[2][0000−0001−9446−3900], Kensuke Tobitani[3][0000−0002−3898−8435], Miyuki Toga[1][0000−0002−2385−4389], and Noriko Nagata[1][0000−0002−2037−1947]

[1] Kwansei Gakuin University, 2–1 Gakuen, Sanda-shi, Hyogo 669–1337 Japan
{k.shinkai,toga.m,nagata}@kwansei.ac.jp
[2] Kobe University, 1–1 Rokkodai-cho, Nada-ku, Kobe-shi, Hyogo 657–0013 Japan
Iori_tani@penguin.kobe-u.ac.jp
[3] University of Nagasaki, 1–1–1 Manabino, Nagayo-cho, Nishi-Sonogi-gun, Nagasaki 851–2195 Japan
tobitani@sun.ac.jp

Abstract. Customized products that can be manufactured according to customers' preferences and applications are attracting attention in the fashion field. One example of a customized product in the fashion field is the "bespoke" suit. Bespoke requires a tailor with expertise. A system that can automatically recommend the impression the user wants would be beneficial because digitizing expert skills would assist less skilled tailors and salespersons. We propose a method for automatically estimating the affective impressions evoked by suit styles using convolutional neural network (CNN) features. The following steps are taken. (1) We quantify visual impressions by subjective evaluation experiments, (2) extract CNN features as physical features, and (3) model the relationship between visual impressions and physical features. We compared three types of CNN features: content features, style features, and features concatenated with content and style features. The correlation coefficients between the evaluation scores and the estimated scores were positive and above moderate for all the features, confirming the effectiveness of the proposed method. The average correlation coefficient of the style feature was slightly higher than that of the content feature, and was not significantly different from that of the feature that concatenated the two features. These results suggest that style features can be effective in estimating not only fabric (texture) but also silhouette impression (shape).

Keywords: Fashion · Suit · CNN · Impression Estimation Model · Kansei · Aesthetic Concepts.

1 Introduction

Advances in the fourth industrial revolution with the development of the Internet have made it possible to meet flexibly the demand for customization according

* Supported by JSTCOI, JPMJCE1314.

to customers' preferences and applications [1]. Especially in the fashion field, customized products are in high demand because different people have different sizes and uses. On the other hand, customized products increase the number of choices for customers, thus forcing them to make many decisions in their purchasing behavior, requiring much effort and time. Therefore, recommender systems that present information in which users may be interested are attracting attention to assist their decision-making. There are two types of recommender systems: collaborative filtering based on purchase history and content-based filtering that is based on similarity of product features. However, collaborative filtering does not allow recommendations to be made without a purchase history, and content-based filtering only recommends similar products. Therefore, there is a need for technology that can estimate individual preferences and needs.

One example of a product that can be customized according to individual preferences and needs is "bespoke" a traditional method of ordering suits [2]. The term "bespoke" refers to the process of tailoring an outfit through a dialogue between a customer and a tailor or a salesperson, and then producing it according to the customer's preferences and demands. Bespoke requires a highly skilled tailor or salesperson with expertise. However, the lack of human resources is a concern. Digitizing bespoke, which requires skilled techniques, is needed to be able to assist a less skilled bespoke tailor and salesperson.

In this paper, we propose a method to model the relationship between visual impressions and physical characteristics of suit styles using convolutional neural network (CNN) features in order to automatically estimate Kansei (affective) impressions evoked by suit styles that consist of texture and shape.

2 Previous Research

Many studies have been conducted on Kansei from the aspect of aesthetic scores and aesthetic values, such as preferences, aesthetic merits, and emotions [3, 4]. However, Kansei also has aspects of aesthetic concepts [5], which are external evaluations such as "flashy" and "slim" that evoke these aesthetic [6, 7]. Such sub-information is essential for accurate prediction of aesthetic scores. In this study, we define them as impressions, and estimate the impressions of suit styles.

Studies have been carried out to model products' physical features and the impressions those products evoke . These include techniques that use color, gloss, surface roughness, and shape as physical features [8], as well as retrieval techniques that use the relationship between features and impressions for 2D images [9]. Tobitani et al. proposed a method for estimating impressions of 3D objects, such as cars, vases, and chairs [10]. The correlation coefficients between the estimation results and the actual evaluation values that people provided were calculated, and a positive correlation was shown, confirming the proposed method's practical effectiveness. However, in this research, only the 3D shape of the product was targeted, not the product's texture.

There have been many studies on texture [11–13]. Sunda et al. used style features as physical features to represent patterned images of clothes, and they constructed models to estimate automatically the visual impressions [14]. Then, they estimated the visual impressions of the test data images based on the constructed models, and they confirmed a strong positive correlation with the impressions that people actually felt. Their research showed that style features were useful in estimating impressions of patterned images.

Considering a suit is a product in which the fabric is made into a body's shape, both the shape and the texture need to be considered when estimating the impression. In proposing a style transformation algorithm, Gatys et al. focused on both content features and style features of images extracted from VGG-19 [15], which is used for generic object recognition [16]. The content features are the feature maps output from the middle layer of VGG-19, and the style features are the Gram matrices of the feature maps. Gatys et al. suggested that content features are features that hold a lot of shape information necessary for generic object recognition, and style features are features that hold a lot of information about colors and patterns in the image. Applying this to suits, content features are strongly related to a silhouette, while style features are strongly related to the fabric. Therefore, in this study, we focus on both content and style features to consider the shape and texture of products, and clarify the relationship between visual impressions and physical features.

3 Proposed Method

In this study, we propose a method to estimate automatically the visual impressions suit styles evoke. The outline is shown in Fig. 1. (1) First, we conduct subjective evaluation experiments to quantify the visual impressions. (2) Next, we extract CNN features, such as content features and style features, as physical features using VGG-19. (3) Then, we model the visual impressions and physical features using random forest to build impression estimation models. Finally, we validate the proposed method's effectiveness by estimating the test data using the constructed impression estimation models.

4 Quantification of Visual Impressions

4.1 Creating Image Data

First, we collected 3,398 images of suits from Deep Fashion [17], a database of fashion images. Next, to create a dataset suitable for impression evaluation experiments and deep learning, we used Deep Image Matting [18] to remove the background and create images of the suit region only. The generated images are shown in Fig. 2.

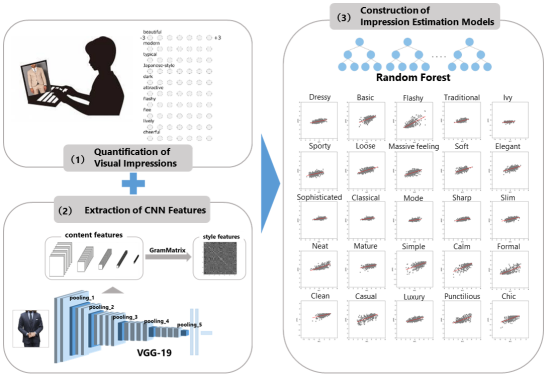

Fig. 1. Overview of our proposed method.

Fig. 2. Image of suit region only.

4.2 Collection and Selection of Evaluation Words

To evaluate properly the suit impressions, it is necessary to collect and select comprehensive and representative evaluation words. It is also possible the words used to describe the impressions suit styles evoke differ between experts and non-experts. Therefore, we collected evaluative words from both experts who have been suit salespersons for 20 years, and non-experts who are undergraduate and graduate students.

First, we collected 26 evaluation words from experts to describe their impressions of suits. Next, we conducted a free-text experiment for non-experts, referring to Tobitani et al. [19] Then, we conducted a goodness-of-fit experiment to verify the obtained words ' suitability for visual impressions. As a result, we collected 26 evaluation words with a high degree of fit. Finally, we conducted a distance measurement experiment to evaluate the substitutability between the evaluation words in semantic space. We did this to select the evaluation words for the impression evaluation experiment from the 52 evaluation words collected from experts and non-experts. After calculating the distance matrix between the evaluation words, we visualized the relationship between the words by using the multi-dimensional scaling and the hierarchical cluster analysis using Ward's method. Based on the results of the cluster analysis, the experts selected 25 evaluation words as those to be used in the impression evaluation experiment. Table 1 shows the selected 25 evaluation words.

Table 1. Twenty-five evaluation words used in the impression evaluation experiment.

dressy	basic	flashy	traditional	ivy
sporty	loose	massive feeling	soft	elegant
sophisticated	classical	mode	sharp	slim
neat	mature	simple	calm	formal
clean	casual	high-class feeling	punctilious	chic

4.3 Impression Evaluation Experiment

To quantify the visual impressions suit styles evoke, we conducted a crowdsourcing impression evaluation experiment on 3,398 collected images. A total of 2240 participants, both experts and non-experts, participated in the experiment. As a result, we obtained 7 levels of evaluation data for each image in 25 different evaluation words for 20 people. In addition, we cleaned the obtained evaluation data, taking into account the evaluation content of the dummy images and the response time. Because of the cleaning, the number of valid respondents was 2,055 out of 2,240 total respondents. Each scale was scored on a scale of 1 to 7 in 1-point increments. We calculated the average score for each rating term and used these scores as the evaluation score for the impression of each image.

5 Modeling the Relationships Between Visual Impressions and Physical Characteristics

5.1 Extraction of CNN Features

As physical features, we used both content and style features proposed that Gatys et al. proposed [16]. We used content and style features output by pooling layers 2, 3, 4, and 5 from VGG-19 [15], which were trained on ImageNet [20] in this study. The dimensionality of each content feature was $56 \times 56 \times 128$, $28 \times 28 \times 256$, $14 \times 14 \times 512$, and $7 \times 7 \times 512$, and the dimensionality of style features was 128×128, 256×256, 512×512, and 512×512.

The dimensionality of these features was too large to be trained, so they needed to be reduced. We applied PCA to the content features and reduced them to make the cumulative contribution ratio 80%. For the style features, considering they are symmetric matrices, the overlapping parts were removed and PCA was applied to reduce them to 80% of the cumulative contribution.

5.2 Random Forest

Random forest is a machine learning method for classification and regression, which uses decision trees to make predictions. In particular, each decision tree does not have high discriminative power, but by combining multiple decision trees, a model with a high accuracy can be constructed. However, considering the structure of a random forest consists of decision trees, the accuracy of the model depends on the tree's depth, and overlearning is likely to occur if the tree

is too deep. Therefore, to set the optimal tree depth, we performed parameter tuning using a grid search. We performed random forest regression in this study, where the objective variables were the evaluation scores and the explanatory variables were the CNN features. For the grid search, the search range of tree depths was set to 3, 5, 8, 10, and 15. We constructed regression models for each adjective in each pooling layer, and the models with the highest correlation coefficient among the models were adopted as the impression estimation models.

Table 2. Correlation coefficients between the evaluation score and the estimated score for each feature.

Evaluation words	Content features	Style features	Content featurs and style features
dressy	0.554	0.565	0.581
basic	0.583	0.626	0.648
flashy	0.573	0.669	0.655
traditional	0.633	0.628	0.642
ivy	0.256	0.396	0.306
sporty	0.679	0.680	0.694
loose	0.679	0.677	0.688
massive feeling	0.641	0.673	0.658
soft	0.621	0.663	0.661
elegant	0.619	0.659	0.648
sophisticated	0.522	0.510	0.500
classical	0.668	0.665	0.664
mode	0.414	0.459	0.427
sharp	0.553	0.568	0.575
slim	0.570	0.580	0.592
neat	0.717	0.701	0.728
mature	0.549	0.585	0.569
simple	0.609	0.633	0.628
calm	0.593	0.663	0.651
formal	0.777	0.765	0.790
clean	0.553	0.595	0.610
casual	0.766	0.748	0.783
high-class feeling	0.637	0.610	0.630
punctilious	0.745	0.733	0.752
chic	0.585	0.625	0.616
average	0.604	0.627	0.628

5.3 Comparison by CNN Features

To verify the effectiveness of the proposed method, we conducted 9-fold cross-validation, comparing three types of CNN features: content features, style features, and features concatenated with content and style features. The correlation coefficients between the evaluation score and the estimated score were used as a measure of estimation accuracy.

The correlation coefficients for each feature are shown in Table 2. The average correlation coefficients of the impression estimation models for all the evaluation words were 0.604 for the case using content features, 0.627 for the case using style features, and 0.628 for the case using features concatenating content and style features, all of which were moderately positive. This indicated our method's effectiveness.

Gatys et al. suggested that content features are features that hold a lot of information of shape, and style features are features that hold a lot of information about colors and patterns. Therefore, in estimating the impression of a suit style consisting of fabric (texture) and silhouette (shape), it is considered that the accuracy can be improved by using features concatenated with content and style features. However, the average of the correlation coefficients for the style features was no significant difference from that for the features concatenated with content and style features, and was slightly higher than that for the content features. This result suggests that the style features retain the information of the content features. In other words, the style features are effective not only for texture impression estimation but also for shape impression estimation. In the future, we plan to conduct a more detailed analysis of the relationship between content features and style features.

6 Conclusion

In this study, we proposed a method for automatically estimating the visual impressions suit styles evoke, and performed the following steps: (1) visual impressions were quantified through subjective evaluation experiments, (2) content features and style features of CNN features were extracted as physical features, and (3) the relationship between visual impressions and physical features was modeled using random forest. We compared three types of CNN features: content features, style features, and features concatenated with content and style features. The impressions on the test data were then estimated based on the obtained models. The correlation coefficients between the evaluation score and the estimated score were more than moderately positive for all features, confirming the proposed method's effectiveness. Comparing the content and style features, the average correlation coefficient of the style feature was slightly higher compared to that of the content feature, and there was no significant difference between the two features when they were concatenated. These results suggest that style features can be effective in estimating the shape impression as well as texture. In the future, we plan to compare features and regression models to construct more accurate models.

References

1. Katou, T.: The new classification of innovation that be recalled with industry 4.0. Journal of International Association of P2M **12**(2), 129–144 (2018)
2. Ross, F.: Refashioning london's bespoke and demi-bespoke tailors: New textiles, technology and design in contemporary menswear. The Journal of The Textile Institute **98**(3), 281–288 (2007)
3. Talebi, H., Milanfar, P.: NIMA: Neural image assessment. IEEE Transactions on Image Processing **27**(8), 3398–4011 (2018)
4. Wang, L., Wang, X., Yamasaki, T., Aizawa, K.: Aspect-Ratio-Preserving Multi-Patch Image Aesthetics Score Prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 0–0 (2019)

5. Sibley, F.: Aesthetic Concepts. The Philosophical Review **68**(4), 421–450 (1959)
6. Black Jr, J. A., Kahol, K., Tripathi, P., Kuchi, P., Panchanathan, S.: Indexing Natural Images for Retrieval Based on Kansei Factors. In Human Vision and Electronic Imaging IX **5292**, 363–375 (2004)
7. Chen, Y. W., Chen, D., Han, X. H., Huang, X.: Generic and Specific Impression Estimation of Clothing Fabric Images Based on Machine Learning. 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), pp. 1753–1757, (2015). https://doi.org/10.1109/FSKD.2015.7382212
8. Niwa, S., Aoyama, Y., Sudo, K., Taniguchi, Y., Kato, T.: Modeling relationship between visual impression of commodities and their graphical features. IPSJ SIG Technical Reports **2013-HCI-152**(24), 1–4 (2013)
9. Chen, Y.W., Huang, X., Chen, D., Han, X.H.: Generic and specific impressions estimation and their application to KANSEI-based clothing fabric image retrieval. International Journal of Pattern Recognition and Artificial Intelligence **32**(10), 1854024 (2018)
10. Tobitani, K., Taguchi, K., Hashimoto, M., Sakashita, K., Tani, I., Hashimoto, S., Katahira, K., Nagata, N.: Impression estimation of 3D object by DNN using multi-view images. The Transactions of the Institute of Electronics, Information and Communication Engineers **J103-D**(11), 844–848 (2020)
11. Julesz, B.: Textons, the elements of texture perception, and their interactions. Nature **290**(5802), 91–97 (1981)
12. Portilla, J., Simoncelli, E. P.: A parametric texture model based on joint statistics of complex wavelet coefficients. International Journal of Computer Vision **40**(1), 49–70 (2000)
13. Tobitani, K., Shiraiwa, A., Katahira, K., Nagata, N., Nikata, K., Arakawa, K.: Modeling of “high-class feeling” on a cosmetic package design. Journal of the Japan Society of Precision Engineering **87**(1), 134–139 (2021)
14. Sunda, N., Tobitani, K., Tani, I., Tani, Y., Nagata, N., Morita, N.: Impression estimation model for clothing patterns using neural style features. HCI International 2020, pp. 689–697, (2020). https://doi.org/10.1007/978-3-030-50732-9_88
15. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. The 3rd International Conference on Learning Representations, (2015)
16. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. The IEEE Conference on Computer Vision and Pattern Recognition 2016, 2414–2423(2016)
17. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. The IEEE Conference on Computer Vision and Pattern Recognition 2016, (2016)
18. Xu, N., Price, B.L., Cohen, S., Huang, T.S.: Deep image matting. The IEEE Conference on Computer Vision and Pattern Recognition 2017, 2970–2979 (2017)
19. Tobitani, K., Matsumoto, T., Tani, Y., Fujii, H., Nagata, N.: Modeling of the relation between I impression and physical characteristics on representation of skin surface quality. The Journal of Image Information and Television Engineers **71**(11), J259–J268 (2017)
20. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., FeiFei, L.: ImageNet large scale visual recognition challenge. International Journal of Computer Vision **115**, 211–252 (2015)