Visualization to Study Bias in Image Annotations and to Support Annotation Based on Semantic Differential

Akari Iijima*
Ochanomizu University

Takayuki Itoh†
Ochanomizu University

1 Introduction

Annotation is a bothering task while creating training datasets for machine learning systems that estimate the impressions of images. This task has another problem that the quality of machine learning depends on the individual differences of annotations of impressions of images. Our study presented in this poster constructs a decision tree that estimates the impressions of images from the impression values provided by many participants so that we can assist the task of annotating impressions of images. This poster proposes a visualization tool for the construction of a decision tree for the annotation of image impressions. In this study, we firstly conducted an impression evaluation applying the semantic differential (SD) method [1], and then generated a fuzzy decision tree [2,3] with the impression values of the images. We can observe the impression estimation process by visualizing the decision tree. Furthermore, we can observe the correlation between the hierarchical structure constructed by the decision tree and the image features by visualizing the classification of images by applying an image browser [4].

2 Impression Evaluation by SD

2.1 Factor selection for the attributes

We applied an image dataset of women's clothes in this study. Based on seral studies on impression analysis of cloth images by researchers in phycology in authors' home country, we adopted the following five factors: color, three-dimensionality, legitimacy, moderateness, and ornamentation, because they are less affected by gender, age, and other individual differences.

Table 1: Impression evaluation data

Image	1500
Participant	43 (20s, 37 female/6 male, 35 JPN/8 CHN)
Attribute 1 'color'	Dark ⇄ Light
Attribute 2 '3-dimensionality'	Fit ⇄ Loose
Attribute 3 'legitimacy'	Formal ⇄ Casual
Attribute 4 'moderateness'	Unusual ⇄ Usual
Attribute 5 'ornamentation'	Simple ⇄ Gorgeous
Likert scale	5

2.2 Pilot testing and attribute determination

Participants may have fatigue if there are a large number of items in the impression evaluation: it may cause the decreased reliability of the impression evaluation. To avoid this problem, we conducted a pilot test to extract appropriate attributes. Here, we prepared 10 clothing images for the pilot test. This test collected responses from the participants (15 Japanese females in their 20s) on the 5-point Likert scale for each evaluation attributes, for the 15 adjective pairs selected from the 5 factors determined in Section 2.1. From the results of this impression evaluation, we calculated the Euclidean distance for actual values obtained in 5-point Likert scale, and then, calculated the sums of distances for each adjective pair. As a result, the adjective pairs that have the smallest sums of distances shown in Table 1 are selected in this study.

2.3 Image pre-processing

This study evaluates the impressions of women's clothing images with a plain, monochromatic background. Since factors other than clothes affect the impression evaluation, we removed the face from the image using OpenCV to detect the facial regions and hide the regions by the background color. Also, the background color of all images is unified by deleting the background. The above process is shown in Figure 1.

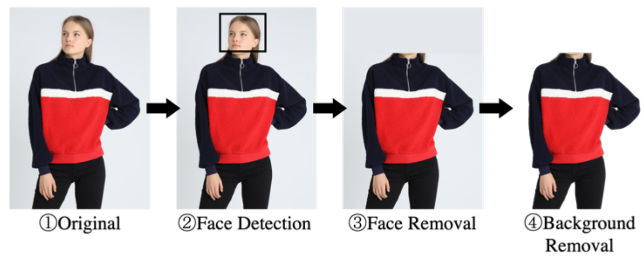

Figure 1: Image pre-processing steps.

2.4 Collection of impression ratings

We collected impression evaluation for 1,500 clothing images from 37 female and six male respondents in their 20s by the five-point Likert scale. They responded with a score on a 5-point Likert scale for each image and each attribute. The number of respondents per image was 12 to 13 because each respondent answered the impression evaluation of 500 images. Table 2 shows the breakdown of the impression evaluations.

3 Visualization Techniques

3.1 Fuzzy decision tree

We applied fuzzy clustering as the clustering method in this study. Fuzzy clustering calculates the attribution to each cluster as a real number in the range [0,1]. The technique enables ambiguous and flexible clustering results. We treat the cluster attribution value as the confidence level of clustering.

The procedure for constructing a fuzzy decision tree is as follows:

- Determine the number of clusters and the depth of the decision tree.
- Calculate the confidence level.
- Construct the decision tree.

We applied "fcmeans" and "iparacoord" modules supported by Python.

Figure 2 shows the procedure for estimating the impressions of "dark-bright" with the five-point Likert scale for a single image. The number of clusters is fixed at 3 ("dark" with the five-point Likert scale for a single image. The number of clusters is fixed at 3 ("dark", "bright", and "neither") and the maximum tree depth is 5 in this implementation. We fixed the maximum tree depth after observing that recursion and repetition appeared more frequently

* e-mail: iijima.akari@is.ocha.ac.jp
† e-mail: itot@is.ocha.ac.jp

when the tree depth was specified to be 6 or more in our experiments.

Then, the procedure generates a decision tree based on the confidence level as shown in Figure 2. We can observe how an image is classified by tracing the tree structure from top to bottom. Each node is drawn as a rectangle and assigned a different color depending on the class type. The histogram shows the number of images that correspond to the confidence level in the range [0, 1]. The bottom node (leaf) represents each class that will depict the clustering result. We can observe the clustering process as well as its confidence level from this visualization. Images whose classes are determined by shallow nodes tend to have high confidence levels, while images whose classes are determined by deep nodes tend to have low confidence levels.

A decision tree is generated for each attribute so that we observe the process of formulating impressions (or annotating) of each image from the visualization results of the decision tree.

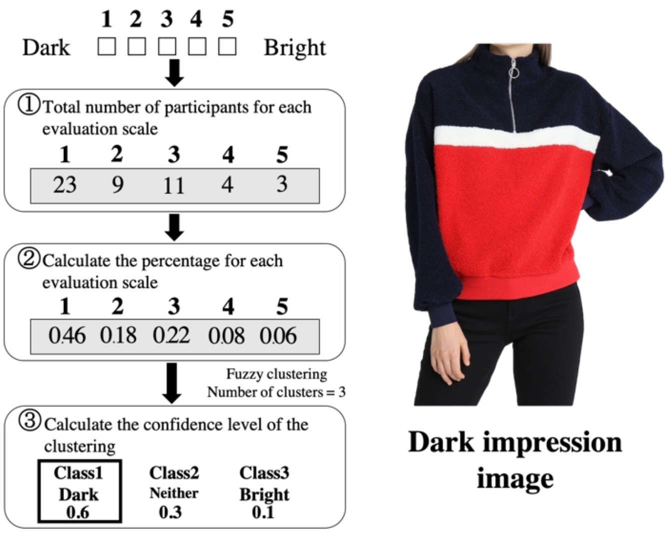

Figure 2: Show the procedure to classify an impression as "dark" from the attribute "dark-bright". The three classes "Dark", "Neither", and "Bright" are selected as the class with the highest confidence level.

3.2 Image Browser

This method generates a frame for each depth of the decision tree and displays it as a dotted line. The frames form a nested structure, where the deepest part of the decision tree corresponds to the inner part in the image browser. The region corresponding to a leaf node of the tree structure is displayed as a group of images.

4 Visualization Example

We implemented this technique by extending the Python visualization library Bokeh. No significant bias was found between the three classes in the clustering result throughout constructing decision trees of the five attributes through the pilot test. We could estimate the appropriateness of the attributes and the strong correlation between the attributes before conducting the impression evaluation. In addition, we could quickly find a group of images that were easy to answer the impression by observing the decision tree. For example, for "casual" in Figure 3, the shallow depth of the decision tree suggests the high confidence among the image groups.

Figure 4 shows an example of an image display in the image browser for "dark-bright". As shown in this figure, brighter impressions are dominated by brightly colored clothing, while darker impressions are dominated by clothing that is almost black. Meanwhile, there are many blue or green clothes, or clothes that combine several colors, for the "neither" impression.

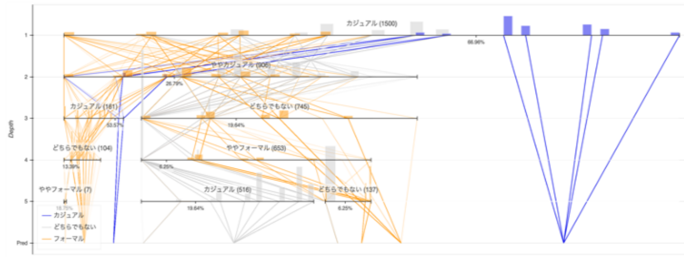

Figure 3: Fuzzy decision tree for the scale "formal-casual". Blue edge is casual, gray edge is neither, orange edge is formal.

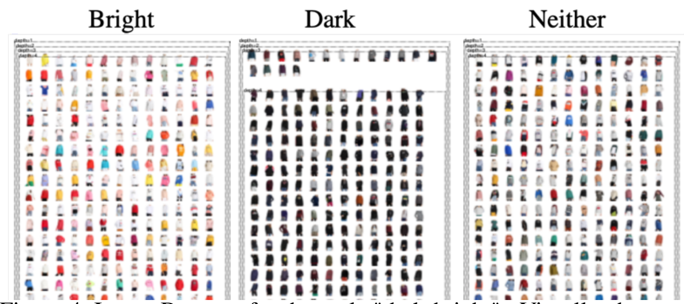

Figure 4: Image Browser for the scale "dark-bright". Visually, the "bright" impression is brightly colored clothing and the "dark" impression is black clothing.

5 Future Work

Our future issues are as follows:

- Removing the bias in impression evaluation
- Improving the readability of decision trees
- Exploring the design of the image browser
- Evaluation measurement of this study

It is important to eliminate the bias of the participants' gender and inputs to improve the reliability and diversity of the data. In addition, we would like to further explore the possibility of knowledge discovery from the image browser by adopting more comprehensive visual representations. After solving these problems, we would like to evaluate our visualization method and discuss its usefulness in terms of semi-automatic annotating.

Acknowledgments

A part of this work was supported by the Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (JSPS). We would like to thank the participants for their cooperation in impression evaluation as annotation.

References

[1] Osgood, C.E., The nature and measurement of meaning. Psychological bulletin, 49(3), p.197, 1952.

[2] Matsuo, Takeshi, et al., Sensitivity Information Analysis of Running Shoes Using Fuzzy Decision Tree and Visualization of Analytical Results. SCIS & ISIS SCIS & ISIS 2010. Japan Society for Fuzzy Theory and Intelligent Informatics, 2010.

[3] Neto, Mário Popolin, Fernando V. Paulovich, Explainable Matrix - Visualization for Global and Local Interpretability of Random Forest Classification Ensembles. arXiv preprint arXiv:2005.04289, 2020.

[4] A. Gomi, R. Miyazaki, T. Itoh, J. Li, CAT: A Hierarchical Image Browser Using a Rectangle Packing Technique, 12th International Conference on Information Visualization (IV08), pp. 82-87, 2008.