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

VIS 2021

Akari Iijima
Ochanomizu University
Tokyo, Japan
iijima.akari@is.ocha.ac.jp

Takayuki Itoh
Ochanomizu University
Tokyo, Japan
itot@is.ocha.ac.jp

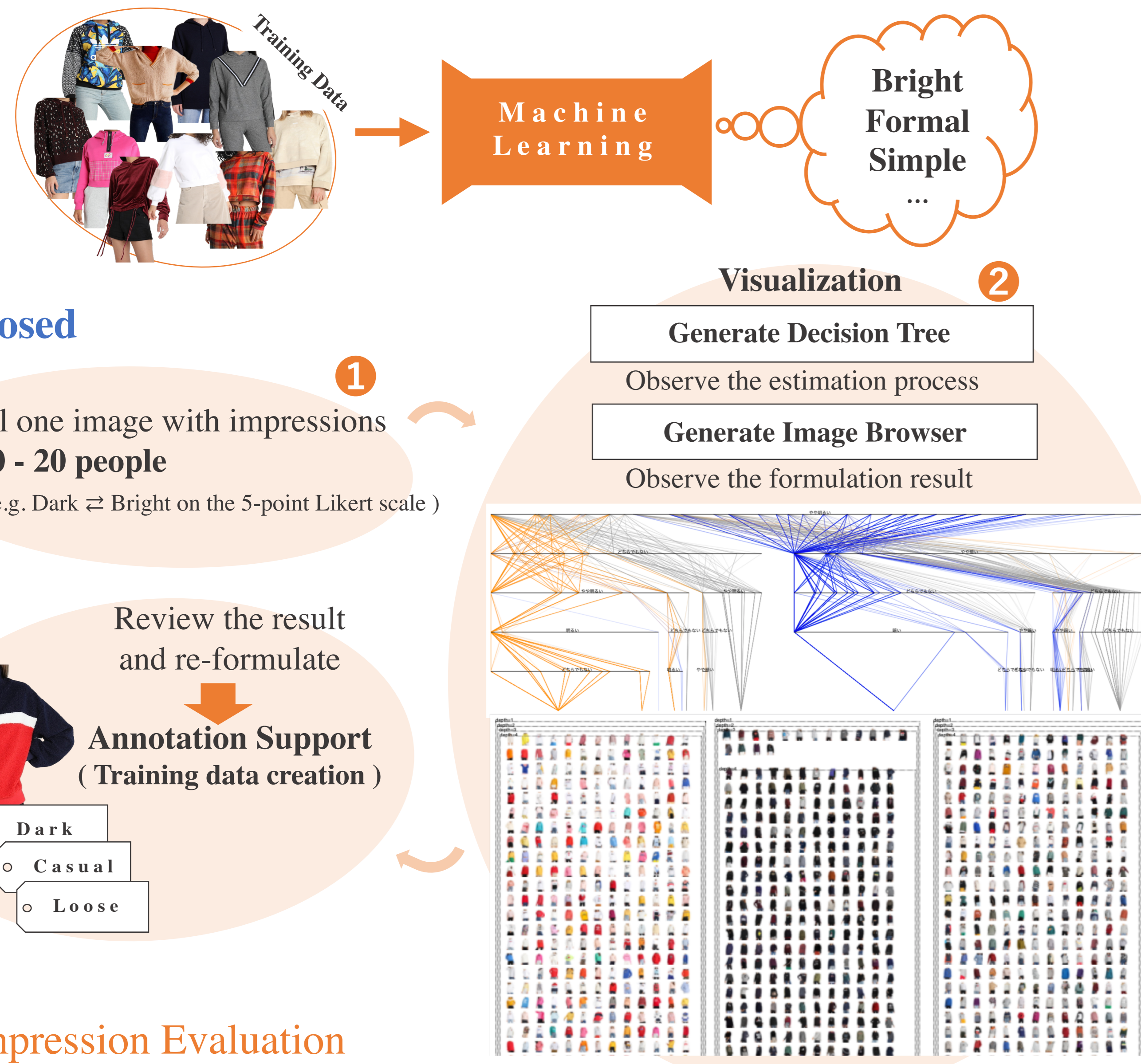IEEE · IEEE COMPUTER SOCIETY · vgtc

Semantic Differential

a rating scale to measure the semantics

Background

- Impression estimation by machine learning
- Training data creation by a single worker annotating many images

Training Data → Machine Learning → Bright Formal Simple ...

Proposed

Visualization ❷
- Generate Decision Tree — Observe the estimation process
- Generate Image Browser — Observe the formulation result

❶
✓ Label one image with impressions by **10 - 20 people**
✓ **SD** (e.g. Dark ⇄ Bright on the 5-point Likert scale)

❸ Review the result and re-formulate → **Annotation Support** (Training data creation)
○ Dark ○ Casual ○ Loose

❶ Impression Evaluation

Factor Selection for the Attributes -

- Five factors that are less affected by gender, age, and other individual differences
- Image dataset of women's clothes

Pilot Test and Attribute Determination -

(To avoid the fatigue of participants)
1. Conduct SD with a larger number of attributes
2. Calculate the Euclidean distance between attributes
3. Select attributes that have smallest sums of distances

Image Pre-processing -

Remove faces and background

① Original ② Face Detection ③ Face Removal ④ Background Removal
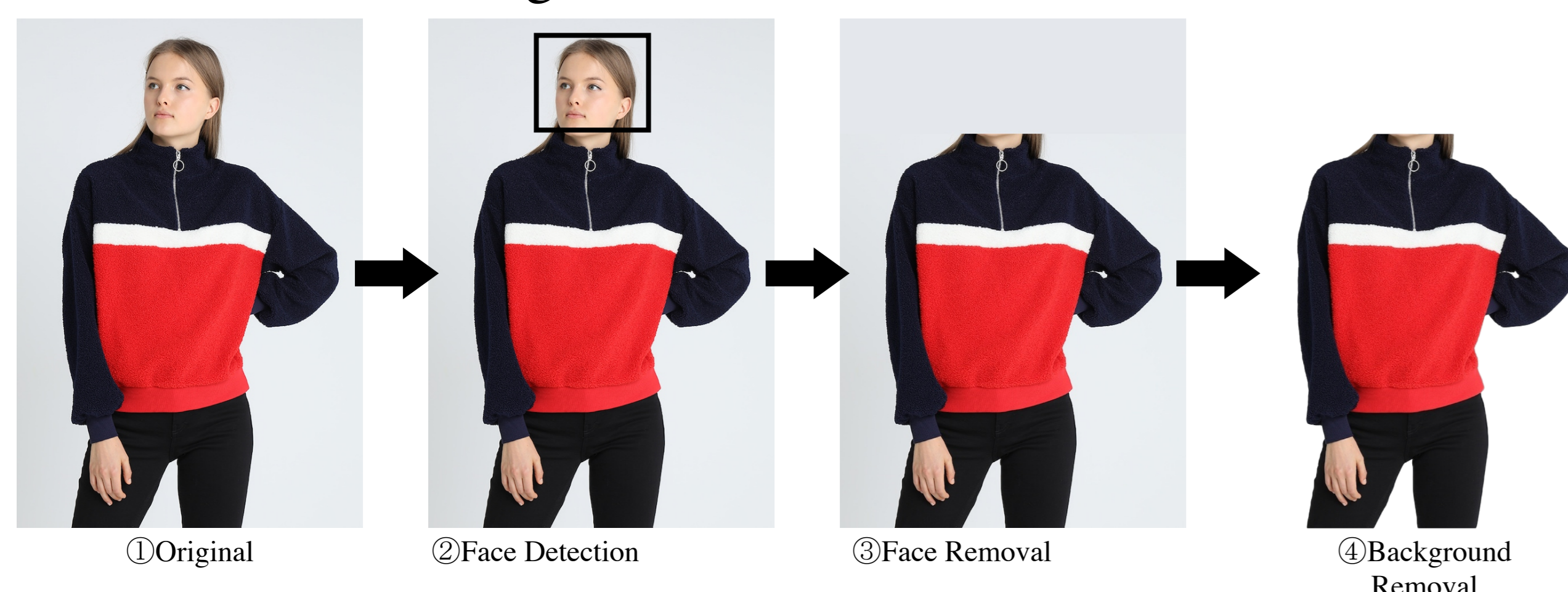
Collection of SD -

Condition: see the table
Environment: participants' own PC and Chrome browser, constant display brightness during the experiment.

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

❷ Visualization

Decision Tree displays how an image is formulated from impression evaluation data by tracing the tree structure

1 2 3 4 5
Dark □□□□□ Bright

① Total number of participants for each evaluation scale
| 1 | 2 | 3 | 4 | 5 |
| 23 | 9 | 11 | 4 | 3 |

② Calculate the percentage for each evaluation scale
| 1 | 2 | 3 | 4 | 5 |
| 0.46 | 0.18 | 0.22 | 0.08 | 0.06 |

Fuzzy clustering
Number of clusters = 3

③ Calculate the confidence level of the clustering

| Class1 Dark 0.6 | Class2 Neither 0.3 | Class3 Bright 0.1 |

How to annotate as Dark impression image

Root	(input) the SD
Leave	(output) formulation results
Branch	image
Node	confidence details
Histogram	the number of images
Color	class (colors have no meaning)

Depth 1 — the higher the confidence level and the easier it is to formulate an impression — Slightly Bright (1500)
Depth 2 — Bright (700) / Neither (800)
Depth 3 — the lower the confidence level and the harder it is to formulate an impression — Dark (400)
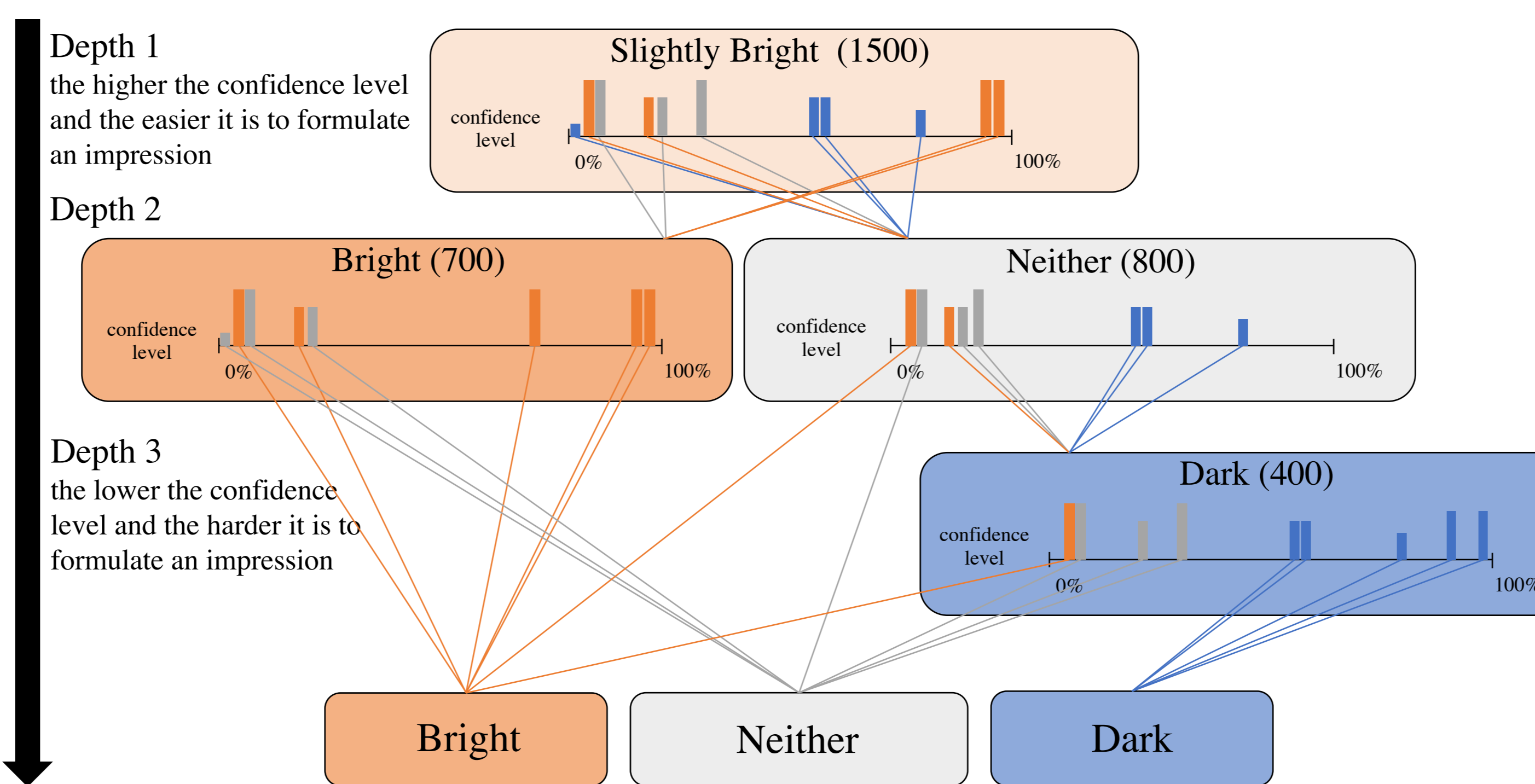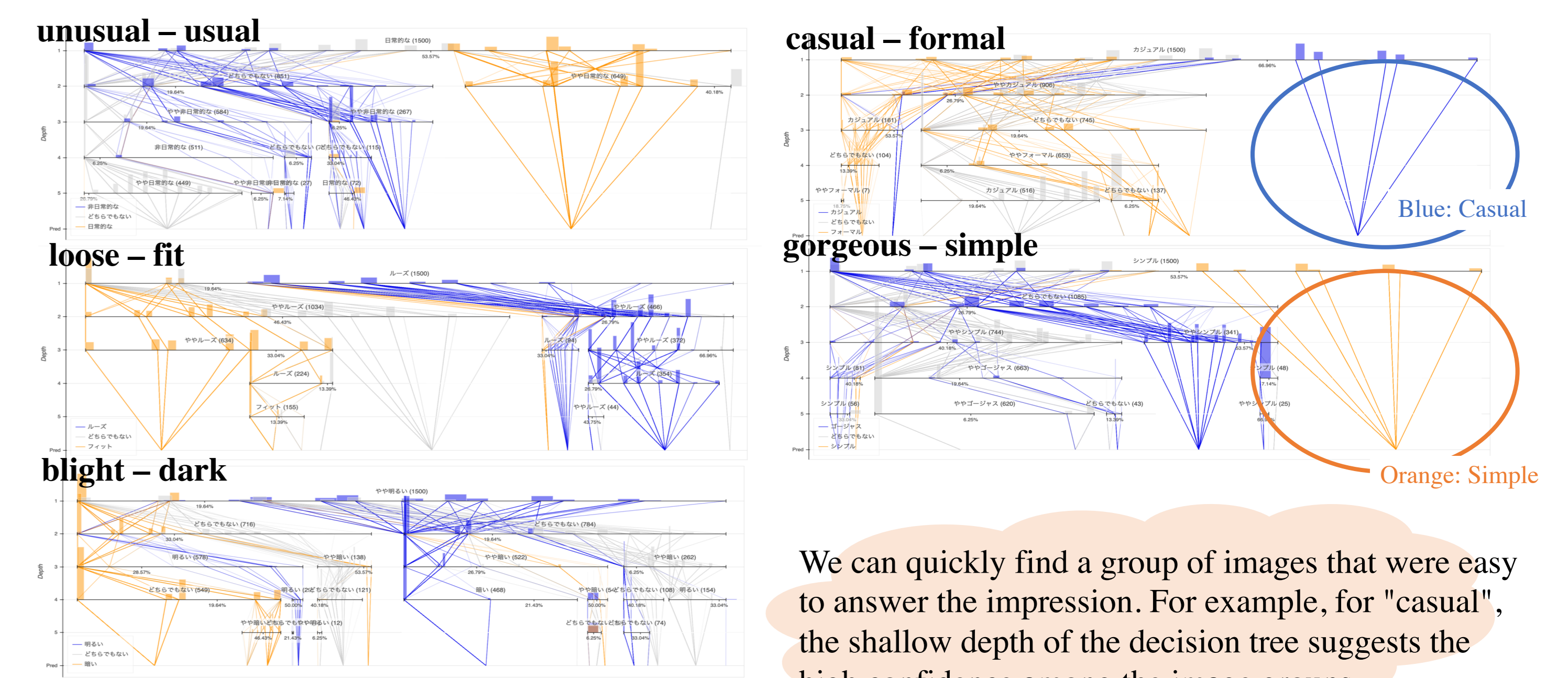
Bright | Neither | Dark

Image Browser shows a frame of the decision tree as a nested structure, where regions corresponding to leaf nodes are displayed as a group of images.

❸ Example

### Example 1 - Decision tree for 5 impression attributes

No significant bias in the clustering result, because of estimation of the appropriateness of the attributes by the pilot test before conducting the SD.

unusual – usual
casual – formal — Blue: Casual
loose – fit
gorgeous – simple — Orange: Simple
blight – dark

We can quickly find a group of images that were easy to answer the impression. For example, for "casual", the shallow depth of the decision tree suggests the high confidence among the image groups.

Example 2 - "dark-bright"

Visually, the "bright" impression is brightly colored clothing and the "dark" impression is black clothing.

Bright | Dark | Neither
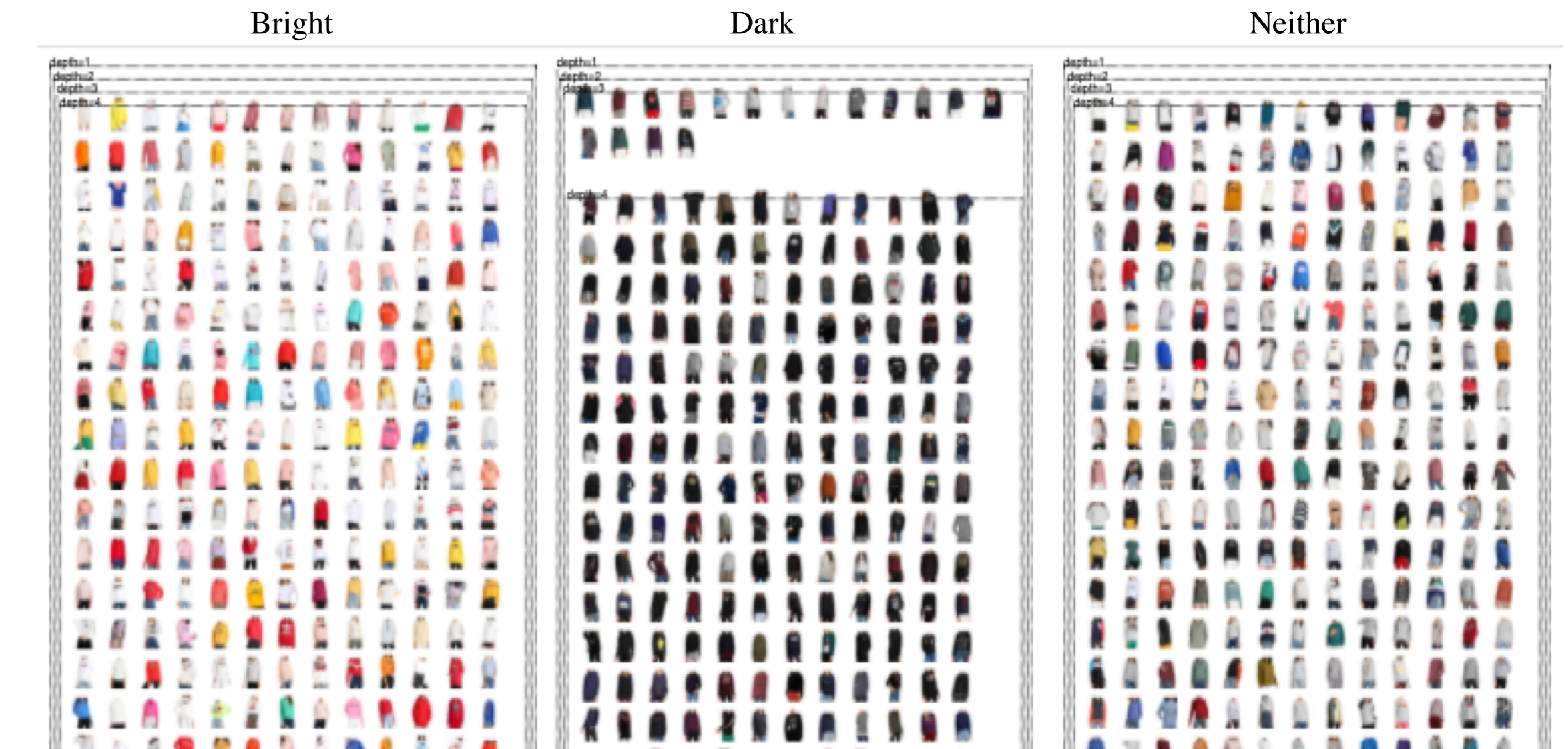
Example 3 -"unusual-usual"

Since more brightly colored the clothes in the impression "unusual", we selected the 4 red clothes images with similar color, shape and design from the contrasting impression "usual" for comparison. It seems unnecessary factors such as pose, hairstyle and physique might affect the impression evaluation.

※ In each attributes , fixed to 3 classes and 5 depth
※ By extending the Python visualization library Bokeh

Unusual | Usual

Future Work

- Removing the input and gender bias
- Improving the readability of decision trees
- Exploring the design of the image browser
- Evaluation measurement of this study