Perceptual Dimensions of Wood Materials

Jiří Filip, A
Jiří Lukavský, B
Filip Děchtěrenko, B
Filipp Schmidt, C
Roland W. Fleming, C

A. The Czech Academy of Science, Institute of Information Theory and Automation
Pod vodárenskou věží 4, 18200 Praha 8
B. The Czech Academy of Science, Institute of Psychology
Hybernská 8, 11000 Praha 1
C. 1. Experimental Psychology, Justus Liebig University of Giessen Germany, 2. Centre for Mind, Brain and Behaviour, Universities of Marburg and Giessen
Otto-Behaghel-Str 10, 35394 Giessen, Germany

Abstract

Materials exhibit an extraordinary range of visual appearances. Characterising and quantifying appearance is important not only for basic research on perceptual mechanisms, but also for computer graphics and a wide range of industrial applications. While methods exist for capturing and representing the optical properties of materials and how they vary across surfaces (Haindl & Filip., 2013), the representations are typically very high-dimensional, and how these representations relate to subjective perceptual impressions of material appearance remains poorly understood. Here, we used a data-driven approach to characterising the perceived appearance characteristics of 30 samples of wood veneer using a ‘visual fingerprint’ that describes each sample as a multidimensional feature vector, with each dimension capturing a different aspect of the appearance. Fifty-six crowd-sourced participants viewed triplets of movies depicting different wood samples as the sample rotated. Their task was to report which of the two match samples was subjectively most similar to the test sample. In another online experiment 45 participants rated ten wood-related appearance characteristics for each of the samples. The results reveal a consistent embedding of the samples across both experiments and a set of 9 perceptual dimensions capturing aspects including the roughness, directionality and spatial scale of the surface patterns. We also showed that a weighted linear combination of eleven image statistics, inspired by the rating characteristics, predicts perceptual dimensions well.

Keywords: texture, surface, colour, categorization, similarity, wood, material, perception, dimension

Introduction

The visual appearance of materials results from a wide range of physical phenomena including the surface's spectral and angular reflectance characteristics, subsurface light scattering, and spatial variations in pigmentation and surface relief. How the visual system estimates such characteristics remains poorly understood (Anderson, 2011, Bracci & Op de Beeck, 2023) and it also remains unclear which perceptual dimensions the visual system uses to describe and compare different materials (Fleming, 2017).

Capturing a comprehensive representation of a surface's physical appearance requires observing it under a sufficient range of illumination and viewing geometries. Complex photorealistic appearances can be approximated by advanced image-based representations used in computer graphics such as the spatially varying bidirectional reflectance distribution function (SVBRDF; Nicodemus & Richmond & Hsia & Ginsburg & Limperis, 1977) or bidirectional texture function (BTF; Dana & van Ginneken & Nayar & Koenderink, 1999). However, these representations are extremely high-dimensional and there is no straightforward mapping between such representations and subjective visual appearance characteristics. Somehow the visual system summarises the overall 'look' of complex, spatially-varying appearances to compare and contrast different materials. Everyday experience suggests that observers do not need to view a material from all possible view- and lighting-directions in order to obtain a distinct impression of its appearance. Yet, although the perceptual representation of materials is surely lower-dimensional than a complete physical description of the surface, there are nevertheless many potential dimensions that the visual system might draw on to describe materials (e.g., overall albedo, relief, glossiness, contrast of surface patterns).

We still do not understand much about such dimensions and how they contribute to observers' judgments of appearance. Which characteristics do observers use to compare different materials? Is there a 'ranking' of characteristics, such that some aspects of appearance dominate comparisons between materials, while others play a secondary role? How specific are certain characteristics to particular classes of materials? Previous work on material perception has often focussed on highly constrained sets of stimuli varying in one or a small number of physical properties (Ferwerda, Pellacini, & Greenberg, 2001; Fleming, Dror, & Adelson, 2003; Fleming, Bülthoff, 2005; Motoyoshi, Nishida, Sharan, & Adelson, 2007; Wendt, Faul, & Mausfeld, 2008; Wendt, Faul, Ekroll, & Mausfeld, 2010; Fleming, Jäkel, & Maloney, 2011; Marlow, Kim, & Anderson, 2012; Paulun, Schmidt, van Assen, & Fleming, 2017; Van Assen, Barla, & Fleming, 2018). Other studies have investigated appearance judgments and categorization based on photographs (e.g., Bell, Upchurch, Snavely, & Bala, 2015; Fleming, Wiebel, & Gegenfurtner 2013; Sharan, Rosenholtz, & Adelson, 2009; Sharan, Liu, Rosenholtz, & Adelson, 2013; Sharan, Rosenholtz, & Adelson, 2014; Wiebel, Valsecchi, & Gegenfurtner, 2013). However, in most cases, it is the experimenters that define which characteristics are judged by participants.

Here we combined this tradition with a more data-driven approach in order to identify dimensions underlying appearance judgments for a set of thirty samples of planar wood veneer with distinctive surface patterns and textures. Wood is a challenging material to characterise due to its complex and varied appearance. It is associated with decorative attributes and is widely used for furniture and interior design. Its structure consists of elongated cells, which are radially oriented rays and longitudinal cells or vessels forming growth rings (Lewin & Goldstein, 1991). Hardwoods tend to have a tighter grain pattern compared to softwoods, resulting in various levels of texture, colour, smoothness, grain density and straightness. All these aspects are impacted by sawing direction and the sample location in the tree trunk. The final visual structure is given by an intersection of a sawing plane with three-dimensional wood structure. Wood has high natural variability in aesthetic characteristics among different species and surface treatments. Previous studies have shown that patterns of anisotropy, colour variations and gloss are the major factors influencing the visual (Nakamura, Masuda, & Shinohara, 1999; Wan, Li, Zhang, Song, & Ke, 2021), multimodal (Fujisaki, Tokita, & Kariya, 2015) aesthetic appeal of wood with impacts on people's preferences (Manuel, Leonhart, Broman, & Becker, 2015), and emotions related to wooden surfaces (Nordvik, Schütte, & Broman, 2009). To the best of our knowledge, all previous studies of wood appearance relied on static stimuli to derive subjective ratings of predefined attributes or their relationship to physical attributes of wood surfaces. This ignores how variable the appearance of even a single sample can be across changes in viewpoint relative to the surface and lighting. Our contribution above this work is twofold.

First, our work uses dynamic (rotating) rather than static stimuli, showing the appearance of the wood samples across variable lighting and viewing conditions. This allowed participants in our experiments to take into account the look of the surface both with and without specular reflections.

Second, instead of relying solely on a possibly incomplete list of predefined visual attributes, we also used similarity judgements to identify the core dimensions underlying judgments of wood. Similarity judgements are an established method for characterising the multidimensional space underlying mental representations, previously used to understand dimensions in object categories (Hebart, Zheng, Pereira, & Baker, 2020), materials (Schmidt, Hebart, & Fleming, 2022) or scenes (Josephs, Hebart, & Konkle, 2023). In contrast to the previous studies, we search for dimensions underlying similarity judgments within a single category.

Specifically, we sought to derive a relatively small number of perceptual dimensions that capture judgments of similarity between movies of the samples. In order to do this, we first crowd-sourced 1218 perceptual similarity judgments from 56 participants. We then applied an analysis method based on sparse, non-negative matrix factorization (Variational Interpretable Concept Embeddings; Muttenthaler, Zheng, McClure, Vandermeulen, Hebart, & Pereira, 2022) to infer a set of dimensions that can predict the similarity judgments. We show that even with a small dataset of thirty samples the method was able to derive visual dimensions that predict the similarity judgments. Specifically, our model identified nine dimensions that together could explain over 75% of the variance in the similarity judgments. Eventually, we showed that standard image statistics obtained from stimuli videos can predict similarity dimensions well.

In addition to the similarity judgments, we also asked a set of 45 participants to judge ten experimenter-defined appearance characteristics for each of the samples (Brightness, Glossiness, Colourfulness, Directionality, Complexity, Contrast, Roughness, Patchiness/regularity, Line elongation, Spatial scale).

The purpose of this was twofold. First, we sought to use the values of these interpretable rating scales to facilitate interpretation of the dimensions derived from the similarity judgments. Second, we sought to cross-validate the embedding of the samples within the 9D space. We reasoned that if different samples are represented in a multidimensional perceptual similarity space—with similar samples close to one another and dissimilar ones further apart—then it should be possible to probe this space through multiple complementary methods (i.e., similarity judgments and subjective feature ratings). We find that the two approaches do indeed lead to similar embeddings of the stimuli, suggesting that they both tap into a common representation within the visual system.

Experiment 1

In the first experiment we collected sparse similarity judgements and used machine learning to infer the full pairwise similarity matrix and to test the embedding of samples in the latent space of wood appearance.

Methods

Participants

Fifty-six participants were recruited using the online crowdsourcing platform Prolific (mean age = 40.5, SD = 16.6, 35 males). All participants reported normal or corrected-to-normal vision and no colour vision impairments. On average, the experiment took 14.0 minutes (SD = 4.9). The participants were reimbursed with 2.1 GBP. All studies within this paper were approved by the Ethics Board of the Institute of Psychology, Academy of Sciences of the Czech Republic (PSU-308/Brno/2022).

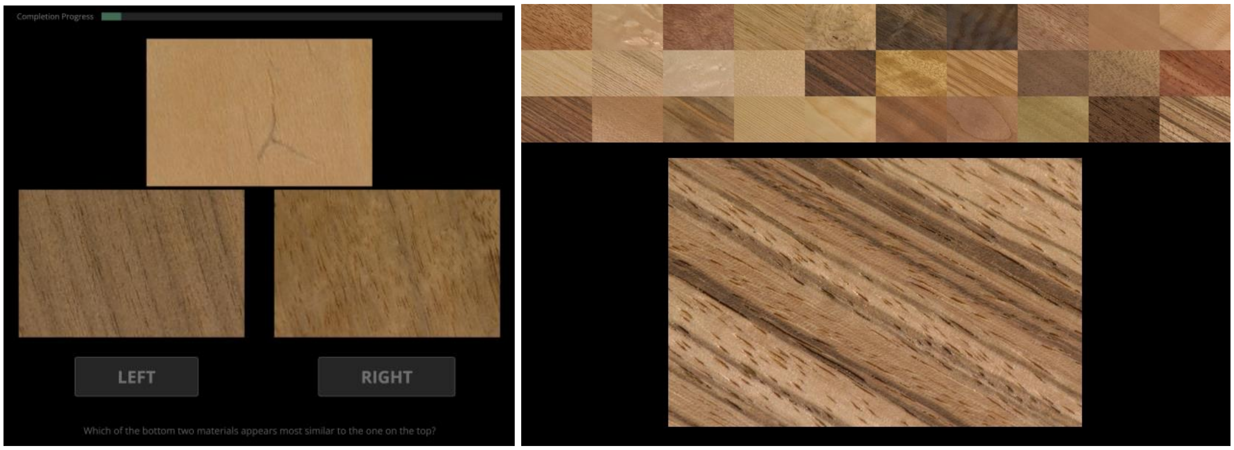
Apparatus and stimuli

We used 30 flat standard wood veneer samples that are used for furniture manufacturing (wood species are listed in Tab.1). We captured video sequences of slow rotations of the samples. Fig. 1 shows the initial (left) and final (right) frame of each video sequence, capturing specular and non-specular view/light geometries. Video samples and additional materials are available at https://osf.io/tz245.

Fig. 1. All 30 samples of wood veneer for (left) specular, (right) non-specular (90 degree rotated) view/light geometries.

All images in the video sequences were 42 x 42 mm areas of the samples, captured by the UTIA goniometer (Filip, Vávra, Haindl, Žid, Krupička, & Havran, 2013). In accordance with industry standards in material observation (McCamy, 1996), we fixed the polar angle of camera and light to 45 degrees and only varied azimuthal angles to allow for faster measurements. Each sequence starts with a difference of 90 degrees between the azimuthal angles of light and camera and includes a movement of the camera by 90 degrees (arriving at a difference of 180 degrees between azimuthal angles), resulting in specular and non-specular material behaviour as shown in Fig. 1. Each 4-second sequence consists of 60 image frames, repeated in reverse order to create a continuous loop of rotating material. See supplementary video [movie_samples_stimuli.avi].

To allow for smooth presentation in the experiment, the image frames of all samples were cropped and downsampled to 400 x 260 pixels, and combined into single-trial frames with three samples on a black background at qHD (quarter high definition, 960 x 540 pixels) as shown in Fig. 2(a). Each sequence was started at a random time of the continuous loop to prevent participants from responding to initial frames of the video sequence.



(a) Experiment 1 (2AFC match-to-sample) (b) Experiment 2 (Ratings)

Fig. 2. Example of stimuli frames of (a) the similarity judgement experiment, where participants responded to: “Which of the bottom two materials appears most similar to the one on the top?”, and of (b) the rating experiment, where participants rated individual samples according to different visual attributes.

Because data was collected online, we did not control for viewing distance (viewing angles) or monitor settings. However, a post-hoc analysis of monitor settings showed a minimal screen resolution of 980 x 577 pixels which allows for a full-resolution presentation of our stimuli.

Experimental procedure

Experiment 1 consisted of 93 trials. In each trial, participants judged the similarity of three presented samples as shown in Fig. 2(a), by deciding which of two match stimuli (at the bottom of the screen) was more similar to the test stimulus (at the top of the screen; 2AFC match-to-sample design). Because we do study similarity within a single material category (wood), we hypothesised a relatively low number of 3 to 5 meaningful perceptual appearance dimensions. In line with the recommendations in (Haghiri, Rubisch, Geirhos, Wichmann, & von Luxburg, 2019) (30 samples and 3-5 dimensions: 900-1500 trials), we tested 1218 triplets, accounting for 10% of the full similarity matrix.

Across all triplets, each sample was presented as a test stimulus in 160-164 trials and as a match stimulus in 304-344 trials. Each triplet was judged four times (i.e, by four different participants). Two out of four repetitions swapped the left and right match stimuli to control for a potential response bias. Each participant was presented with one of 28 unique trial sets or its copy with swapped match stimuli.

Data were collected online using a custom script in the jsPsych framework (De Leeuw, 2015). After reading the instructions, participants completed three practice trials and 90 experimental trials (87 trials plus 3 catch trials). They initiated each trial by clicking the “Start” button after which a video with the three samples started looping (Fig. 2(a)). Participants responded to the instruction below the video (“*Which of the bottom two materials appears most similar to the one on the top?*”) by clicking on the “LEFT” or ”RIGHT” button at the bottom. The response stopped the loop and initiated the next trial, with a progress bar at the top showing the number of remaining trials. Catch trials were presented at fixed positions (40th, 65th, and 84th trials) and featured the same sample presented twice, as standard and match stimulus, yielding a ground truth correct response.

Data analysis

All data is available from the following public repository: [link provided upon acceptance]. We next sought to identify a set of perceptual dimensions—with values for every sample—that could account for the observed pattern of similarity responses. To do this, we analysed the responses using Variational Interpretable Concept Embeddings (VICE; Muttenthaler, Zheng, McClure, Vandermeulen, Hebart, & Pereira, 2022). This algorithm takes as input the sparse (i.e., incomplete) similarity matrix obtained in the similarity rating experiment and estimates the full pairwise similarity matrix. In the process, it iteratively estimates a set of underlying dimensions that could account for the observed responses. As our similarity judgement study comprises 2AFC task, we applied target matching instead of odd-one-out procedure.

Several of the VICE algorithm's hyperparameters can affect its results, including the number of dimensions. To validate the performance of the model, we created random splits of our participants' similarity judgements into training (90% of responses) and test sets (10% of responses). Then, we performed a limited grid search for selected hyperparameters of the model: learning rate [0.0005, 0.001, 0.002], mixture of distributions in the spike-and-slab prior [Gaussians, Laplace], spike (a prior of probability at zero values) [0.125, 0.25, 0.75], slab (a prior of probability for the non-zero values) [0.2, 0.5, 1.0], and probability of relative weighting of the distributions [0.4, 0.5, 0.6]. The training typically converged within 200 epochs, and typically resulted in between 8 and 11 dimensions (min. 4, max. 14 dimensions). Details of the model selection and training process are reported in Section 1 of the supplementary material.

Results

Consistency of similarity judgement responses

Our results show that participants were highly consistent in their similarity judgments. When analysing inter-individual consistency based on the four repetitions of each triplet, in 569 triplets (47%) all four responses were the same, in 439 triplets (36%) three responses were the same, and in 210 triplets (17%) responses were on par. This suggests that in the majority of trials (87%) subjects were consistent, only in the remaining 17% they were at chance. Also, when comparing sequences with their copies with swapped match stimuli, only in 61 trials (5%) swapping resulted in a different response.

Deriving perceptual dimensions from similarity ratings

Based on the parameter grid search (see Section 1 of the supplementary material), we picked the best performing model (9 dimensions; accuracy on the training set = 0.760; accuracy on the test set = 0.769). Importantly, even though the number of dimensions varied between different resulting models from the parameter grid search, the meaning of those dimensions was highly preserved. Specifically, the embeddings obtained from the first five best VICE models (with 4-9 dimensions) were highly similar (mean correlation between similarity matrices of the 4 next best models to that of the best model was R=0.939). Thus, in the following we analyse the best performing VICE model under the justified assumption that it is representative of a family of models with similar embedding.

The resulting embedding as shown in Fig. 3(a) is quite sparse, with on average only 6 values > 20% percentile in each similarity dimension. Fig. 3(b) shows the sum of loadings for individual dimensions and suggests that the first 5 dimensions have higher impact than the remaining 4. Fig. 3(c) compares how well the similarity responses from participants can be approximated by the values estimated from the VICE model. Chance performance in the 2AFC match-to-sample task (red) is 50%, with the inter-participant noise ceiling (grey) at 82%. The noise ceiling is computed as the average consistency across the four repetitions of each triplet, and represents the best possible prediction any model could achieve for our dataset, given the variation in the data.

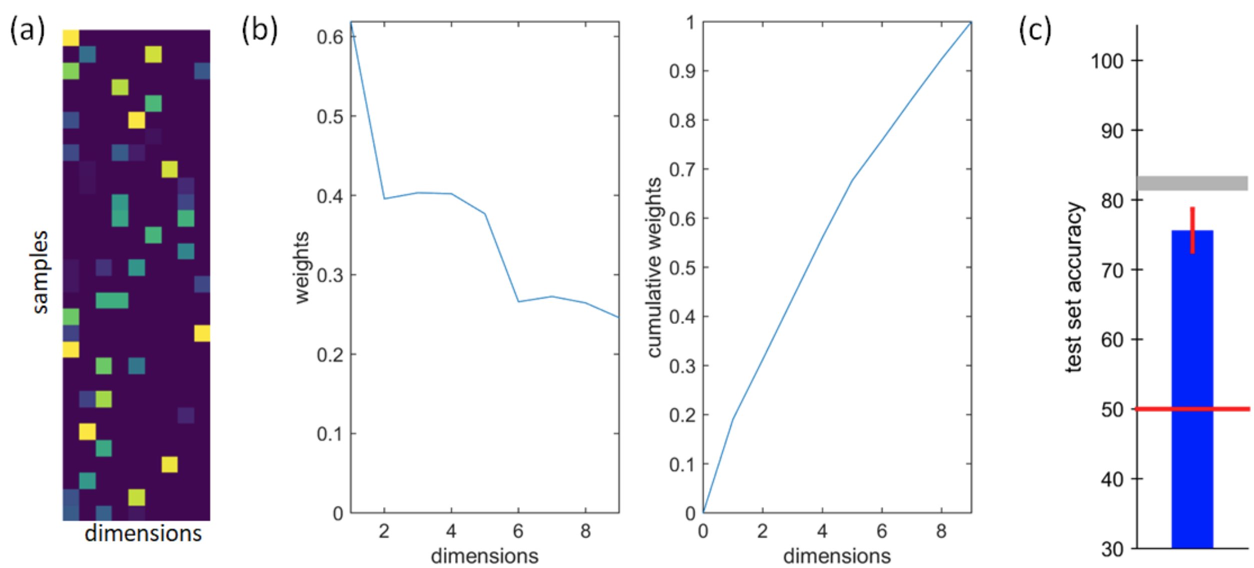


Fig. 3 Details on the 9 dimensions of the best VICE model: (a) estimated embedding, (b) dimensions loadings, and (c) average accuracy on test set (blue) with 95% confidence interval error-bar (red), noise ceiling (grey), and chance level (red).

Fig. 4 shows samples rank-ordered by their embedding values in each of the 9 dimensions (highest values to the left). Each video sample is represented by its two most distinct frames, i.e. non-specular and specular reflection (refer to the supplementary to see the dynamic behaviour of the actual video samples).

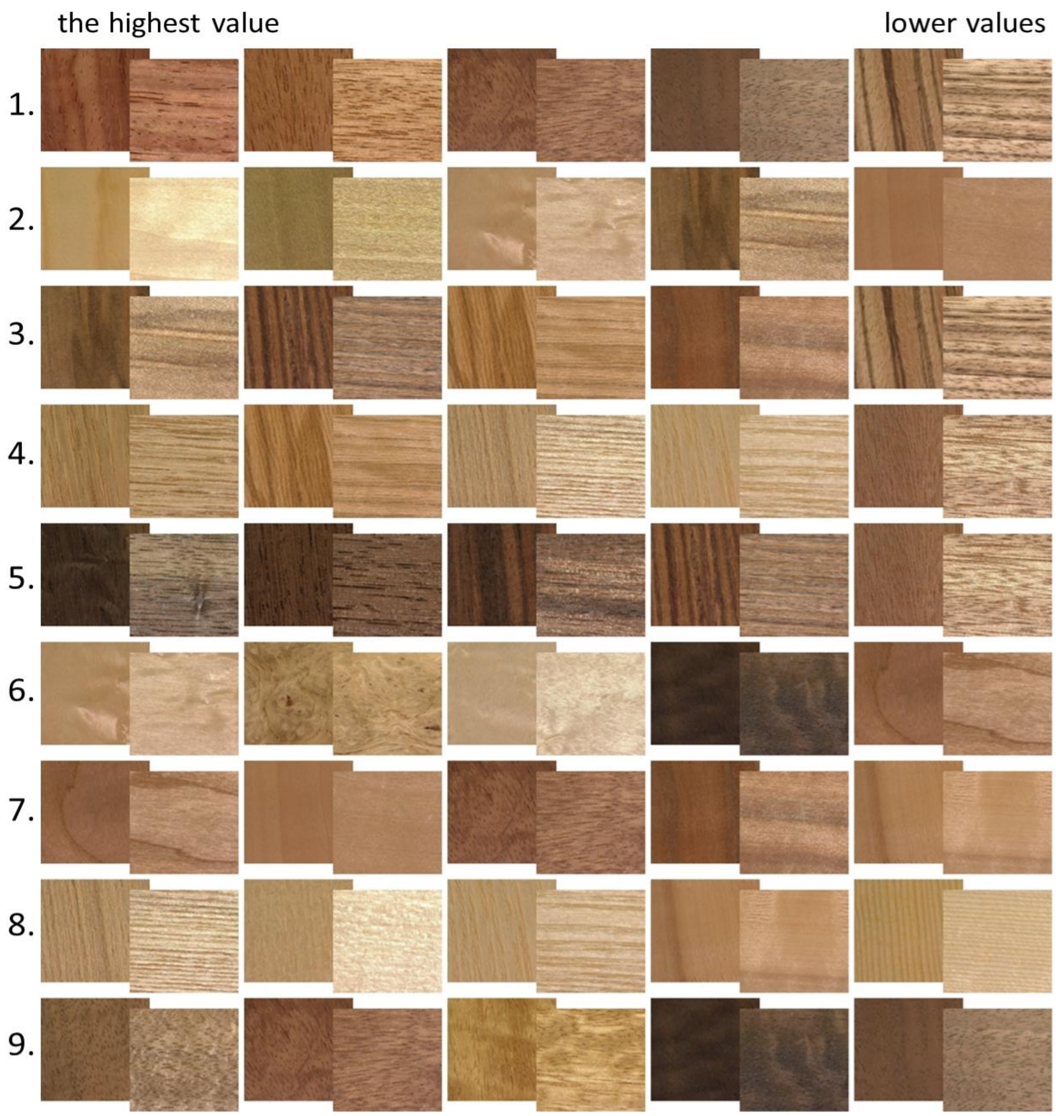


Fig. 4 Five samples for each dimension rank-ordered based on embedding values. Each video sample is represented by both the most non-specular and most specular condition. See left side of the supplementary video [movie_similarity_vs_rating.avi].

The full pairwise similarity matrix of the wood samples that we obtained from the estimated embedding is shown in Fig. 5. We used hierarchical clustering (based on weighted average Euclidean distance) to cluster similar samples together, showing that samples had approximately three main visual modes, which might be visually interpreted as rough/contrast (M1), spatial frequency (M2), and directional (M3). These modes are present also in individual similarity dimensions in Fig. 4, where M1 is represented by dimensions 1, 5 and 9, M2 by 2 and 6, and M3 by 4, 8 and 3. Note that similar modes were also found using Louvain community detection method (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008) as reported in Section 2 of the supplementary material.

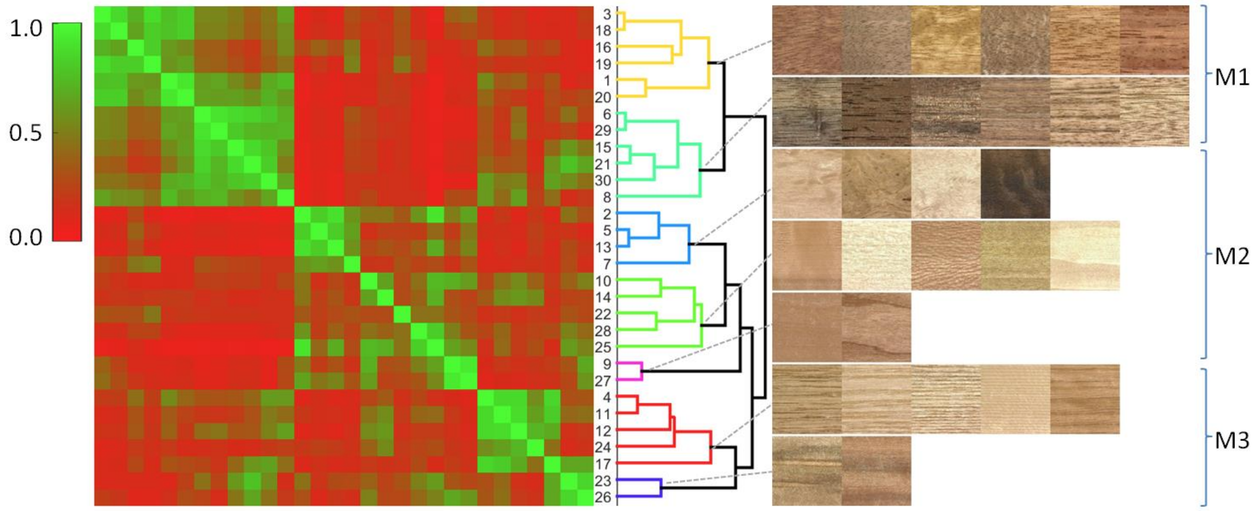


Fig. 5 Estimated pairwise similarity matrix with samples ordered based on hierarchical clustering, and the depiction of the corresponding samples in the individual clusters.

Discussion

The analysis of participants' similarity judgements using the VICE model provided us with 9 visual appearance dimensions of wood. However, even though visualising the embedding by ranking samples within each dimension may provide some intuition about the meaning of the dimensions, it is not clear whether these intuitions are the best description of the respective dimensions. For this reason we performed a second comparative experiment relying on standard attributes rating on a Likert scale.

Experiment 2

The main goal of the second experiment was to obtain perceptual judgements for all wood samples for a set of visual appearance attributes widely used in the field of material perception. By being able to describe our samples in terms of these specific perceptual attributes, we aimed to provide a more valid interpretation of the similarity dimensions from the first experiment–and a corresponding understanding of the main visual cues that naive observers use to describe and discriminate between types of wood.

Methods

Participants

Forty five volunteer observers participated in the online experiment (age data were not collected). All participants reported normal or corrected-to-normal vision and no colour vision impairments. On average, the experiment took 22.0 minutes (SD = 17.6).

Apparatus and stimuli

The stimuli used in Experiment 2 were the same as in Experiment 1.

Procedure

Participants were presented with 30 trials, each showing one of the sample videos from Experiment 1. The resolution of each stimuli image was 920 x 600 pixels. To make the task easier for participants, all other materials were simultaneously presented for comparison at a smaller scale at the top of the screen as shown in Fig. 2(b). Participants rated each material on ten visual appearance attributes (brightness, glossiness, colourfulness, directionality, complexity, contrast, roughness, patchiness/regularity, line elongation, and spatial scale), using a visual analog scale. The attributes were selected based on a review of previous research (Tamura, Mori, & Yamawaki, 1978; Rao & Lohse, 1996; Fleming, Wiebel, & Gegenfurtner 2013; Tanaka & Horiuchi, 2015, Nordvik, Schütte, & Broman, 2009) and salient differences between samples identified by the experimenters. For the participants, the meaning of each visual attribute was explained with a short sentence (e.g., brightness: "*How bright is the material in comparison with the others?*"). Also, the end points of each scale were labelled (e.g., brightness: "dark" and "bright"). A full description of each visual attribute and the corresponding endpoint labels is provided in Section 4 of the Supplementary Material.

All attribute scales were on the screen simultaneously, and at the start of each trial all sliders were set to the centre of each scale. Only after moving all sliders, participants could proceed to the next trial.

Data analysis

Again, a post-hoc analysis of monitor settings showed a sufficient minimum screen resolution of 980 x 768 pixels. The inter-rater agreement was determined using intraclass correlation coefficient (ICC; Koo & Li, 2016, with two-way random effects, based on mean rating and consistency). More detailed analysis of participants' responses is provided in Section 5 of the Supplementary Material.

Results

The rating responses for each attribute formed unimodal distributions with mean values close to the central point (45.8 to 59.5) and similar SD values (21.7 to 29.6). The ICC indicated excellent reliability (ICC > 0.898) for all attributes but *spatial scale* where ICC = 0.659 indicated only moderate reliability.

Fig. 6 Five samples for each rating dimension rank-ordered based on average rating responses. Each video sample is represented by both the most non-specular and most specular condition. See right side of supplementary video [movie_similarity_vs_rating.avi].

Samples with the highest rating responses for each rating dimension are shown in Fig. 6, with visually intuitive results in the majority of dimensions (again with the exception of *spatial scale*).

Note that these examples also suggest similarities between rating dimensions (i.e. overlap in samples for e.g. *colorfulness* and *contrast*). To measure these inter-class similarities, we computed Pearson correlations for mean rating values across all 30 samples. As shown in Fig. 7, we observe a high similarity between *colorfulness-contrast, directionality-line elongations* and *complexity-patchiness/regularity*. On the other hand, a high dissimilarity is observed for *brightness-colorfulness* and *brightness-contrast*. These similarities are also evident at the level of individual samples, as is shown in Fig. 10(a) which is showing similarity matrices for individual rating dimensions.

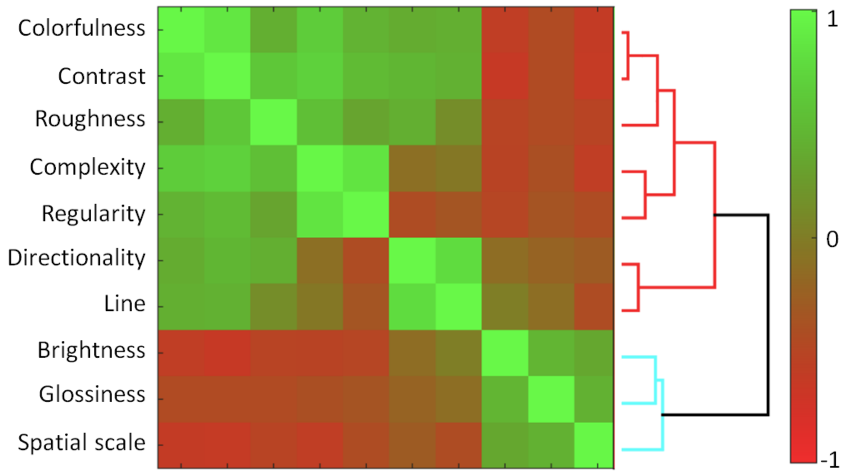


Fig. 7 Inter-class similarity, computed as Pearson correlation across all samples, with the dendrogram showing the results of hierarchical clustering of attributes.

See supplementary video http://staff.utia.cas.cz/filip/tmp/movie_similarity_vs_rating.avi with material samples ranking as a function of dimensions loadings of VICE (left) and rating responses having the highest and the lowest values.

Discussion

Our rating experiment provided reliable and visually intuitive data on the selected visual appearance attributes, but also highlighted mutual dependencies between some of the attributes. This suggests that our samples can be described by less than 10 attributes, that is, the latent visual dimensionality of our samples is lower than 10. In the next section, we compare the visual dimensions obtained from the similarity and rating experiments.

Interpretation of similarity dimensions

As the meaning of the similarity dimensions discovered by the VICE model are not known, we used cross correlation and multilinear regressions between appearance ratings and similarity judgements as well as between their respective similarity matrices. This allowed us to assign meaning to the similarity dimensions by relating them to the meaningful appearance ratings.

Cross-correlation of similarity and rating dimensions

Across all appearance attributes, the correlation between similarity matrices from ratings and similarity judgements is relatively low (Pearson R=0.335; exclusion of matrix diagonal), with the highest correlations for *directionality* (R=0.448) and *contrast* (R=0.462). This confirmed our expectation that the similarity embedding cannot be explained using a single rating dimension.

For a direct correlation between all ratings and all similarity dimensions see Fig. 8(a). The highest positive correlation was R=0.744 and the highest negative correlation was R=-0.813. Notably, similarity dimensions 1, 3, 4 and 5 show similar patterns of correlation to rating attributes *colourfulness*, *directionality*, *complexity* and *roughness*. On the other hand, similarity dimension 7 is not correlated strongly with any rating attribute, which suggests that none of them can explain the visual appearance captured by this particular dimension. To test whether the similar pattern of correlations across similarity dimensions follows from a strong dependency between individual rating attributes, we computed PCA on our rating data. Fig. 8(b) shows that only four PCA components explain 91.1% of the variance, suggesting that the effective number of main visual appearance dimensions for our set of wood samples is about 5-10. We confirm this hypothesis by using a statistical approach to estimate the number of dimensions based on triplet embedding accuracy of ordinal triplets embedding (Künstle, von Luxburg, & Wichmann, 2022) – which identifies 6 as the inherent dimensionality of our data (see details on this analysis in Section 3 of the supplementary material). This is also supported by the steep drop of similarity embedding factor loadings with more than five dimensions (Fig. 3(b)).

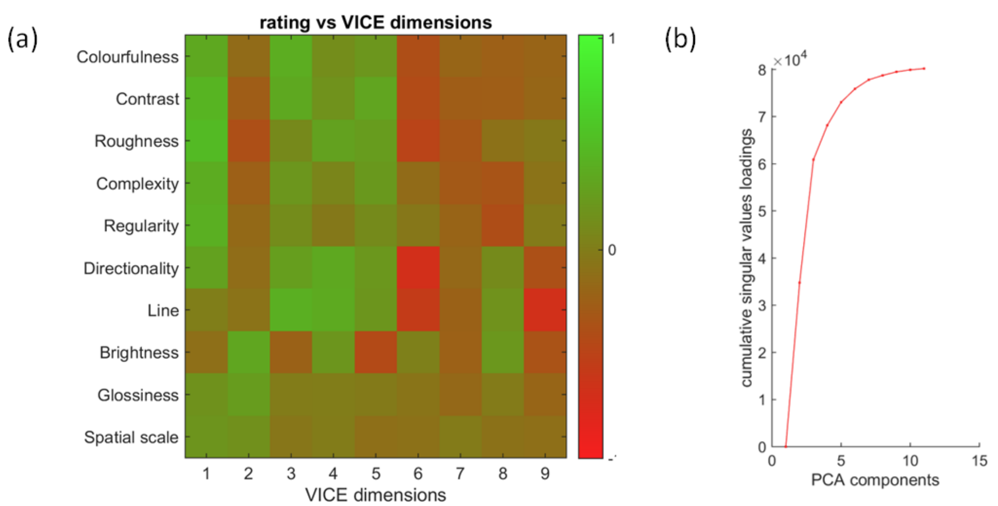


Fig. 8 (a) Correlations of rating dimensions (rows) to similarity dimensions (columns), with negative correlations in red and positive correlations in green (range [-1,1]). (b) Cumulative singular values loadings of PCA computed on correlations across rating dimensions (a).

A more quantitative comparison between similarity dimensions and rating attributes is shown in Fig. 9. For each similarity dimension, we ordered and scaled samples according to their dimension values. The inset shows how well the variation in each similarity dimension is correlated with different rating attributes. Here we observe similar patterns for dimensions 1, 3, 4, and 5 while dimension 7 is virtually constant across rating attributes. R^2 scores in the legend demonstrate how well each similarity dimension can be predicted by a linear regression of rating attributes.

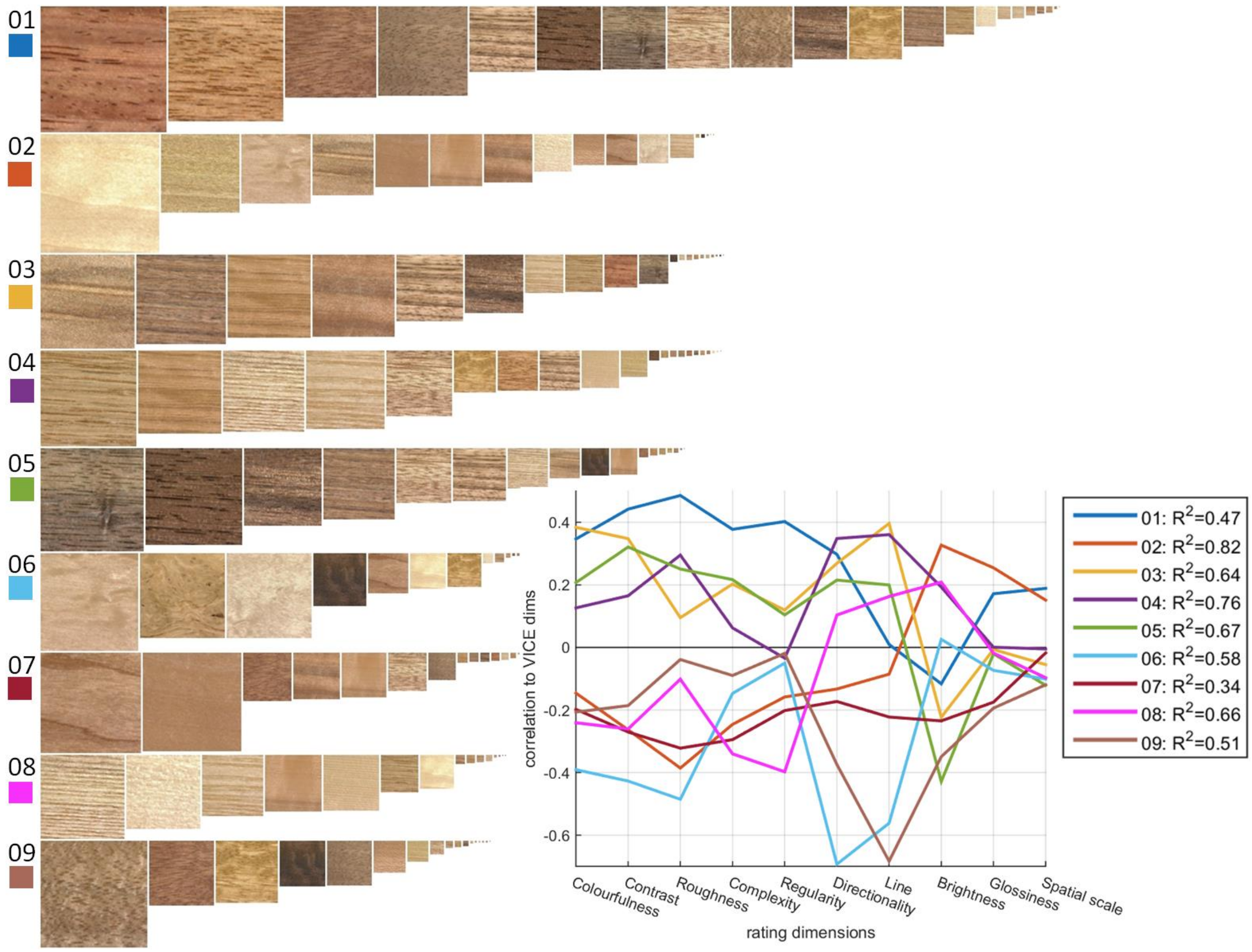


Fig. 9 Sample rank-ordered by embedding values in VICE similarity dimensions. Inset: Correlations between similarity (VICE) and rating attributes, obtained from linear regressions (R^2 scores provide information on how well the linear regression using rating dimensions explained individual similarity dimensions. See the supplementary video [movie_similarity_scaled.avi].

To evaluate similarity of results obtained from both experiments, we computed the rating similarity matrix as Euclidean distance across all attributes. A direct correlation between similarity matrices obtained from similarity judgement and attributes rating (excluding diagonal elements) was R=0.627 (R^2 = 0.393). The matrices are shown in the first row of Fig. 10(a,b).

To assess the main visual dimensions for similarity judgements, we computed multidimensional scaling MDS (Carroll & Arabie, 1998) on the VICE similarity matrix. The MDS projection of samples onto the first two dimensions are shown in Fig. 10(d). In line with our visual interpretation of the three main visual modes in Fig. 5, the first MDS dimension can be interpreted as related to roughness, the second to directionality and the third to spatial frequency. For clarity, we also included these plots with the video samples as presented to observers. We compared MDS results over the similarity matrices and coordinates of all 30 samples for the first two MDS dimensions after Procrustes alignment are shown in

Fig. 10(e). For MDS of VICE similarity matrix into all 3 dimensions see a top part of the supplementary video [movie_MDS_simmat_linreg.avi].

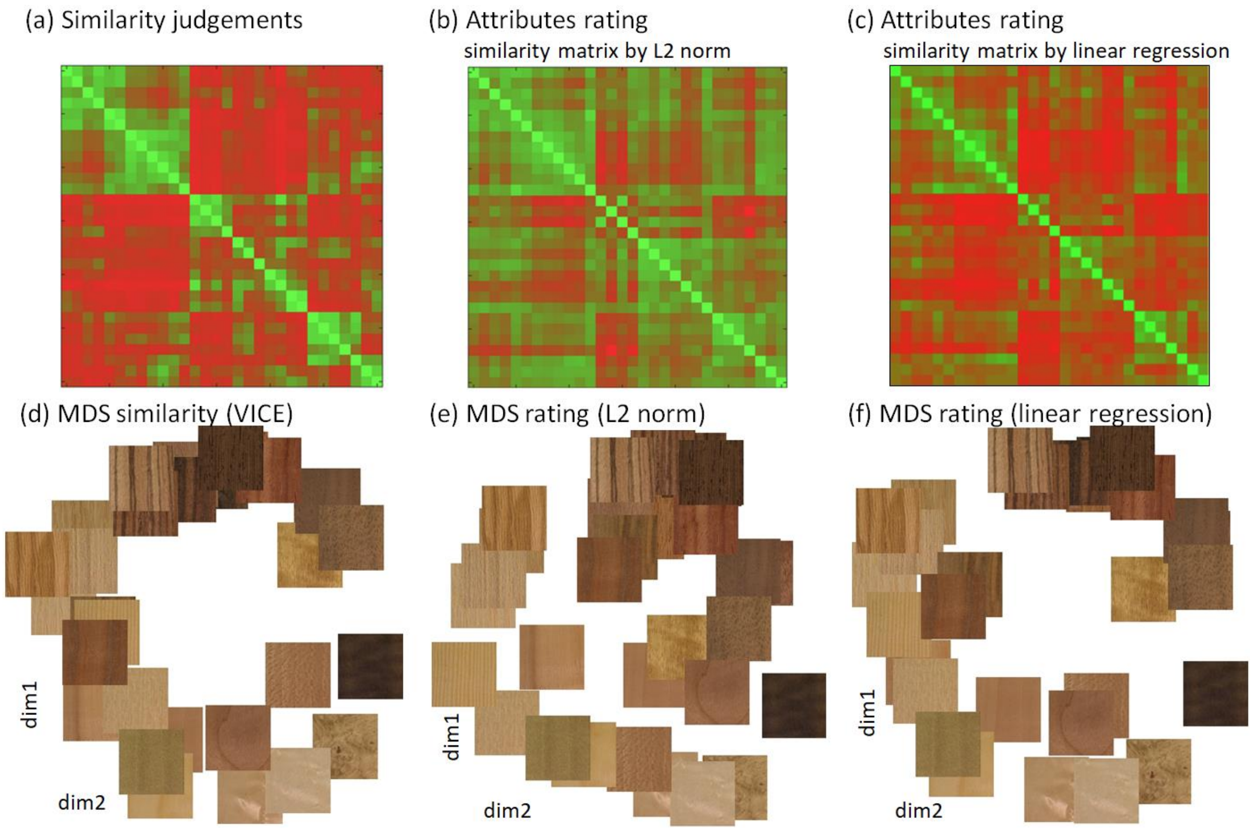


Fig. 10 A comparison of similarity matrices obtained by (a) similarity judgements and (b,c) ratings using L2-norm and linear regression, respectively. The correlation between matrices is for (b) R=0.627 (R^2 = 0.393) and for (c) R=0.723 (R^2=0.523). Corresponding embeddings of samples in the first two MDS dimensions for (d) similarity judgement and (e,f) ratings (after Procrustes alignment).

Prediction of similarity matrix from rating attributes

Beyond simple correlations between individual ratings and similarity dimensions, we can test how well a combination of rating attributes predict similarity judgements. To this end, we used multilinear regression to predict the similarity judgement matrix by a linear combination of the rating attribute similarity matrices shown in Fig. 11(a). The matrices' diagonals were kept to anchor scaling. The regression model explains about 52% of the variance in similarity judgements (R=0.723, R^2=0.523), while still preserving the major similarity modes as shown in Fig. 11(a). To evaluate the importance of individual rating attributes for the reconstruction, we performed leave-one-out regressions and the resulting drops in explained variance. Fig. 11(b) shows that the most important attributes are *brightness*, *directionality*, and *roughness*. A comparison of the obtained multi-dimensional scaling over the similarity matrices and coordinates of all 30 samples for the first two MDS dimensions after Procrustes alignment are shown in Fig. 10(f). Also see a Section 6 of the supplementary material for samples alignment according to MDS and video [movie_MDS_simmat_linreg.avi], comparing three MDS dimensions of similarity judgements similarity matrix (top) with its linear regression using rating attributes (bottom).

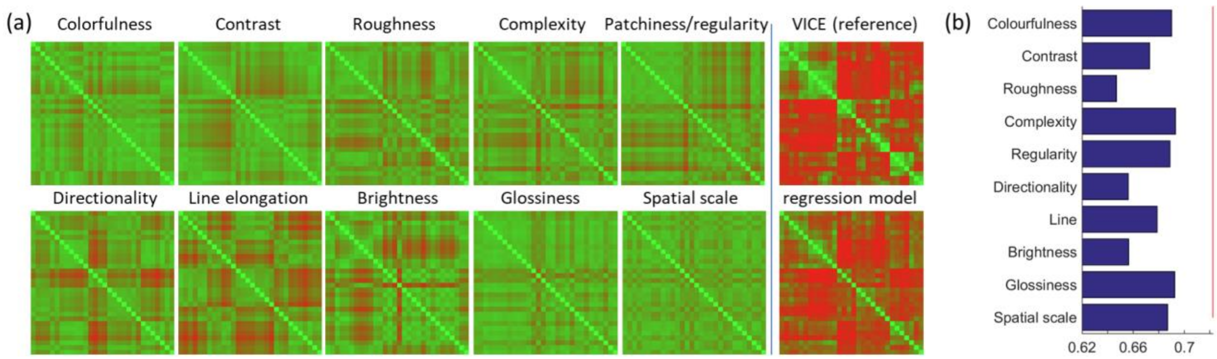


Fig. 11 (a) Similarity matrices of individual rating attributes compared to the VICE similarity matrix and the result of the linear combination of the 10 rating similarity matrices. (b) Results of leave-one-out regression analyses showing the respective drops in correlation below the red baseline due to individual attributes removal.

Prediction of similarity dimensions from rating attributes

Finally, we used linear regression to predict individual similarity dimensions by a linear combination of the rating attributes. Across all dimensions, ratings can well explain similarity dimensions, with an average of R=0.851 (R^2=0.731). R^2 scores of similarity dimensions represented by the regression model are shown in Fig. 12(a). All dimensions except 7 and 9 can be well explained by a combination of rating attributes. As reported previously, dimension 7 is not well predicted by any of the rating attributes. This might be for two reasons: either none of our predefined attributes is not capturing the same visual appearance as that dimension, or there is a general bias in our rating data that is introduced by a particular interpretation of the to-be-rated attributes. For instance *line elongation*, *patchiness/regularity* or *spatial scale* might have different meanings at different frequency scales. For instance samples 22 and 30 (see Fig. 1) both share a fine detail structure and a distinct low-frequency stripy pattern. As a result, observers might be confused as to whether these attributes should be evaluated on a fine or coarse scale, resulting in overall ambiguous ratings. In Fig. 12(b), we are plotting normalised regression values to visualise the contribution of rating attributes to each of the similarity dimensions. For example, dimension 1 is strongly negatively related to colorfulness and roughness, but strongly positively related to contrast, regularity and directionality; while dimension 5 represents materials that were judged low on line elongation, brightness and glossiness (see samples rank ordering along dimensions in Fig. 9).

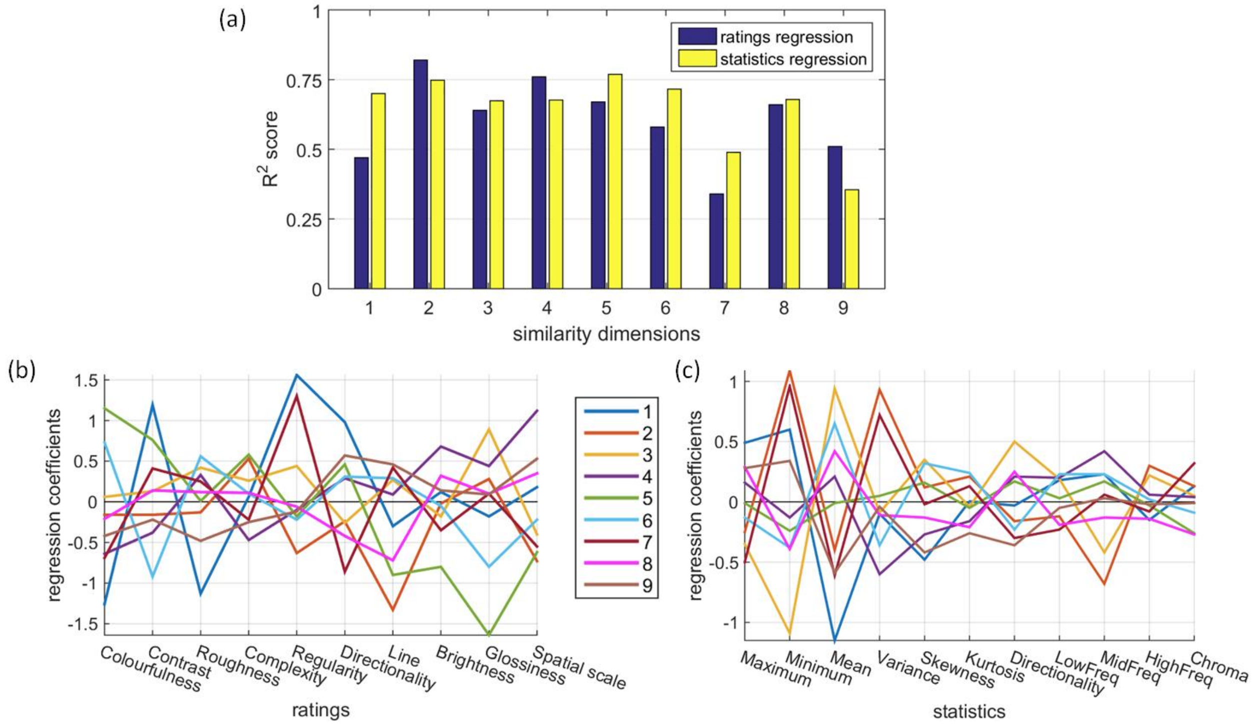


Fig. 12 (a) R^2 scores of similarity dimensions regression of rating dimensions (blue) and regression of computational image statistics, with corresponding normalised regression coefficients showing contribution of rating dimensions (b) and computational statistics (c) for reconstruction of each similarity dimension.

Relationship to computational statistics

To relate similarity dimensions to computational statistics, we used standard image statistics used in texture synthesis related to human low-level perception of textures (Portilla & Simoncelli, 2000, Motoyoshi, Nishida, Sharan, & Adelson, 2007), namely *minimum, maximum, mean, variance, skewness,* and *kurtosis*. We supplied additional statistics evaluating image directionality (Maskey & Newman, 2021) and frequency content in three bands (*low, mid*, and *high* frequencies) computed from PSD of image converted to the Fourier domain. The final values of statistics were averaged across all frames of movie sequence. We used these statistics for linear regression of similarity ratings and R^2 scores of results are shown as yellow bars in Fig. 12(a). We observe similar values of R^2 scores to those obtained from rating regression and in general all similarity dimensions, except 7 and 9, can be represented reasonably well using our statistics. Mean R^2 score across all dimensions was 0.65 (R=0.73). Normalised regression coefficients of individual statistics are shown in Fig. 12(c). For example dimension 1 has the highest coefficients for *maximum* and *minimum*, which relates to contrast, while dimension 2 has the highest coefficients for *minimum* and *variance* which relates to spatial variations within the structure as we can observe in typical representants in respective dimensions in Fig. 9. Also see a supplementary video [movie_MDS_simmat_stats.avi], comparing three MDS dimensions of similarity judgements similarity matrix (top) with a similarity matrix obtained as Euclidean distance of all eleven statistics (bottom).

General discussion

In this study, we set out to identify the perceptual core characteristics of wood. Characterising the visual appearance of wood is complex because of the variety in factors like colour, grain patterns, fine-scale relief and reflectance behaviour. Accordingly, a description in physical terms requires very high-dimensional measurements that capture the image projected by the material surface across all possible lighting conditions and viewing angles. Yet, we reasoned that when human observers are asked to compare samples—or judge the appearance of a single sample—they would rely on a relatively small number of dimensions that together summarise the overall 'look' of each surface and its texture—what we might call a 'visual signature' of the material (Sharan, Liu, Rosenholtz, & Adelson, 2013; Schmidt, Hebart, & Fleming, 2022).

Here, we wanted to estimate such an internal multidimensional representation by asking observers to make comparisons between samples. A secondary goal was to test the extent to which different methods of probing this putative representation yielded similar embeddings of the material samples. We reasoned that if observers draw on shared, core perceptual dimensions to judge the appearance of wood, it should be possible to probe this representation using distinct tasks.

To test this, we performed two experiments using movies of thirty samples of different wood veneers, rotating in such a way as to reveal both non-specular and specular appearance modes. In the first experiment, we took a data-driven approach, asking participants to make relative similarity judgments in a 2AFC task, from which we sought to derive underlying dimensions using the VICE algorithm (Muttenthaler, Zheng, McClure, Vandermeulen, Hebart, & Pereira, 2022). In the second experiment, we defined a set of ten appearance characteristics and asked participants to rate each sample in terms of all ten characteristics, effectively directly stating the location of each sample in a ten-dimensional appearance space. Our main findings can be summarised as follows:

- In Experiment 1, the VICE algorithm revealed that nine dimensions could account for 75% of the variance in the similarity judgments, consistent with the notion of a low-dimensional 'visual fingerprint' summary representation of their appearance.
- In Experiment 2, participants were consistent in their judgments of the ten appearance characteristics, suggesting agreement about the embedding of samples relative to one another.
- Comparisons between the two experiments showed a significant overlap between embeddings of the samples derived from the two tasks, providing further evidence for a core representation of wood materials, with similar-looking samples close to one another, and more distinct ones further away from one another within the multidimensional appearance space.
- The consistency between the two experiments can also be demonstrated by approximating the dimensions inferred from Experiment 1 as a weighted linear combination of the ratings in Experiment 2.
- Finally, a set of quite simple low-level image features, designed to capture similar appearance characteristics as the rating dimensions predict the ratings and VICE dimensions surprisingly well, using simple linear regression. Although these image features will not be the exact quantities that the visual system uses to represent and compare the wood samples, this shows how we can use straightforward image-computable models to predict perceived differences in

appearance (under constant viewing conditions). This has potential practical applications in many areas.

Our study also provides a proof-of-principle demonstration that it is possible to establish embeddings of items from a single basic-level category (here: wood) within a perceptual space using either a subset of all possible similarity comparisons, or through direct rating of particular features. The study differed from previous investigations in the use of movies rather than static images, capturing a wide range of appearances for each sample, and in the comparison between similarity and appearance ratings.

Limitations and future directions

Although our study provides a first proof-of-principle for identifying perceptual dimensions within categories, there are a number of important limitations of the approach, which we consider here.

Limited number of wooden samples

The stimulus set considered here consisted of only thirty samples of different wood veneers, as listed in Table 1. This is one of the largest sets of wooden samples used in a psychophysical analysis to date, and we carefully selected this set from a catalogue of over one hundred wood veneers so as to provide as broad and uniform a range of appearances as possible. However, including a larger number of samples would necessarily provide additional information about the embedding, and would potentially reveal additional perceptual dimensions by covering a wider range of appearances. It would also be particularly interesting to include in future work multiple samples of each species (see Table 1), to capture within-item variability as well. We would expect that although different samples would be clearly discriminable, generally they would tend to occupy very close locations within the multidimensional perceptual space.

Limited observation and illumination geometry

By using dynamic stimuli, in contrast to previous studies, which tended to offer only a single view of each sample, we were able to provide observers with some information about how the appearance of the samples changed depending on viewing conditions, including both specular and non-specular conditions. Nevertheless, this still represented a limited subset of all possible lighting-sample-viewer configurations. We had to limit camera and light trajectories so that movies were of reasonable duration. Based on pilot work with a range of different sampling parameters, we identified a rotation that was of acceptable durations and that was intuitive for observers. As the appearance of wood does not typically change much with polar angle, we limited polar viewing angles to 45^{o} and changed azimuthal angles only. A comparison of image histograms from our videos with those of the full BTF for the same material (at polar angles 45^{o} including over 400 images for different combinations of illumination and view azimuthal angles) provided mean differences of X^2 lower than 0.10. This leaves us confident that the selected views were representative of the overall appearance.

Table 1 A complete list of wood species used in the experiment.

01	afzelia	11	white ash	21	rosewood
02	masur birch	12	ash heartwood	22	plane
03	pommele bubinga	13	maple burl	23	satinwood
04	oak	14	European lime (linden)	24	spruce
05	burr oak	15	macassar ebony	25	spruce knotted
06	smoked oak	16	movingui (lemon)	26	tineo
07	eucalyptus	17	olive	27	American cherry
08	gaboon	18	European walnut	28	tulipwood
09	pear	19	Peruvian walnut	29	wenge
10	European apple	20	padauk	30	zebrawood

Limited size of samples

On a related point, the visible area of the samples was around 50x50mm. This size was selected to deliver fine surface details. On the other hand, for certain species, there may be low-frequency content that was excluded by the small size. To compensate for this during video acquisition the location of the captured area on the veneer specimen was carefully selected to demonstrate the main sample's characteristics. A similar comparison of histogram statistics with BTF data over a large scale of image plane resulted in similarly low differences in histograms, again indicating that the patch was representative of the sample as a whole.

Limited coverage of triplets for similarity judgements

In Experiment 1, we measured only a small subset of all possible stimulus triplets. Specifically, our experiment had a coverage of 10%, which is nevertheless far greater than the less than 2% coverage used in other studies using related data analyses (Hebart, Zheng, Pereira, & Baker, 2020). On the other hand, our number of samples is considerably lower, greatly reducing the number of necessary trials. We followed the recommendations in (Haghiri, Wichmann, & von Luxburg, 2020) to estimate the number of judgements, although future studies could potentially increase the coverage further for small stimulus sets like ours.

Stability of dimensions

Statistical inference methods like VICE are stochastic, so repeated runs of the algorithm on the same data can deliver slightly different outcomes. This naturally raises questions about the stability and interpretation of the outcome. We tested a wide range of hyperparameter values, and found the values we used delivered representative results. Importantly, although the exact number of dimensions varied across runs, the meanings of those dimensions (i.e., the loadings across samples) were highly conserved. This, along with the high extent to which the dimensions could predict similarity ratings gives high confidence that the analysis delivered robust results. Increasing the number and diversity of samples, as well as the coverage would lead to even greater stability, although with obvious practical costs. It is nevertheless important to emphasise that in interpreting results on small and constrained stimulus sets like ours, greater emphasis should be placed on the *embedding of items* within the multidimensional space than on the precise number or direction of the dimensions returned by VICE (or related algorithms). The convergence between the ratings and the VICE analysis supports this view.

Intuitive interpretability of individual dimensions

While some studies (e.g., Hebart, Zheng, Pereira, & Baker, 2020; Josephs, Hebart, & Konkle, 2023; Schmidt, Hebart, & Fleming, 2022) have found that analyses similar to VICE deliver dimensions that are highly intuitively interpretable, in our case, most of the dimensions appeared to be better understood as weighted combinations of more intuitive factors. This can be seen in Fig. 9, for example, in which samples are ranked by their values of the nine dimensions returned by VICE. Some of the dimensions seem to capture intuitive concepts. For example, dimensions 4 appears related to stripiness, and this is consistent with the high loading of the 'Directionality' and 'Line' features in the multiple regression for this feature. Dimension 6, in contrast, seems to be approximately the opposite, with an emphasis on samples with turbulent texture patterns rather than linear grain. However, for most of the other dimensions the interpretation is less intuitive. This is likely due to the small and constrained sample set. With diverse image sets that span the entire range of commonly occurring objects, for example (Hebart, Zheng, Pereira, & Baker, 2020), almost all samples will have near-zero values of any given attribute, while there are still sufficient numbers of images with high values to enable a dimension to emerge from the analysis. Indeed, such datasets are particularly well suited for seemingly meaningful individual dimensions to be recovered by the sparse nonnegative matrix factorization. By contrast, within-category samples, as in our experiments, tend to involve characteristics that are more uniformly distributed across samples. This is likely to be one of the reasons that the recovered dimensions were composites of multiple factors. Nevertheless, again it should be noted that we place greater emphasis on the embedding of items within the space than on the exact orientation of the underlying dimensions.

Choice of rating dimensions

There are practical limits to the number of appearance attributes that participants can feasibly be asked to rate for each sample. As with the majority of previous perceptual studies of wood surfaces (Nakamura, Masuda, & Shinohara, 1999; Nordvik, Schütte, & Broman, 2009; Fujisaki, Tokita, & Kariya, 2015; Manuel, Leonhart, Broman, & Becker, 2015, Wan, Li, Zhang, Song, & Ke, 2021) we preselected a list of visual properties in our rating experiment. This list, of course, is likely to be incomplete as there are potentially infinitely many ways of describing samples, including those that may make intuitive visual sense, but which cannot easily be put into words. Nevertheless, we find that this set of dimensions leads to intuitive and repeatable judgments, which are sufficient to capture an embedding of the samples similar to that revealed by the similarity ratings and VICE analysis. Future studies could also ask participants, rather than the experimenters, to provide terms that describe important appearance differences between samples, which other participants would then rate (see, e.g., Van Assen, Barla, & Fleming, 2018).

Conclusions

Our study sought to identify core perceptual dimensions underlying the appearance of wood. Using thirty movies of rotating planar wooden veneer samples, we asked participants to judge the similarity between items and rate each sample along ten predefined dimensions. The results revealed a consistent embedding of samples between the two tasks, suggesting a core internal representation of the samples, capturing the overall 'look' of the samples in a relatively small number of dimensions. These could be expressed as a weighted linear combination of the following ten attributes: brightness, glossiness, colourfulness, directionality, complexity, contrast, roughness, patchiness/regularity, line elongation, and spatial scale. The results not only reveal the core dimensions underlying the perception of wood, they also provide a proof of concept demonstration for how perceptual dimensions underlying judgments within a single basic-level category can be probed using multiple tasks.

Acknowledgements

This research has been supported by the Czech Science Foundation grant GA22-17529S as well as the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation—project number 222641018—SFB/TRR 135 TP C1), and by the Research Cluster 'The Adaptive Mind', funded by the Excellence Program of the Hessian Ministry of Higher Education, Science, Research and Art.

References

Anderson, B. L. (2011). Visual perception of materials and surfaces. *Current biology, 21*(24), R978-R983.

Bell, S., Upchurch, P., Snavely, N., & Bala, K. (2015). Material recognition in the wild with the materials in context database. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3479-3487).

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *2008*(10), P10008.

Bracci, S., & Op de Beeck, H.P. (2023). Understanding Human Object Vision: A Picture is Worth a Thousand Representations. *Annual Review of Psychology, 74*, pp.113-135.

Carroll, J. D., & Arabie, P. (1998). Multidimensional scaling. *Measurement, judgment and decision making*, 179-250.

Dana, K.J., van Ginneken, B., Nayar, S.K., & Koenderink, J.J. (1999). Reflectance and texture of real-world surfaces*, ACM Transactions on Graphics*, Vol.18, Issue 18, pp.1-34

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior research methods*, *47*, 1-12.

Ferwerda, J. A., Pellacini, F., & Greenberg, D. P. (2001). Psychophysically based model of surface gloss perception. In *SPIE Human vision and electronic imaging vi,* Vol. 4299, pp. 291-301.

Filip, J., Vavra, R., Haindl, M., Zid, P., Krupicka, M., & Havran, V. (2013). BRDF slices: Accurate adaptive anisotropic appearance acquisition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1468-1473).

Fleming, R. W., Dror, R. O., & Adelson, E. H. (2003). Real-world illumination and the perception of surface reflectance properties. *Journal of vision, 3*(5), 3-3.

Fleming, R. W., & Bülthoff, H. H. (2005). Low-level image cues in the perception of translucent materials. *ACM Transactions on Applied Perception, 2*(3), 346-382.

Fleming, R. W., Jäkel, F., & Maloney, L. T. (2011). Visual perception of thick transparent materials. *Psychological science, 22*(6), 812-820.

Fleming, R. W., Wiebel, C., & Gegenfurtner, K. (2013). Perceptual qualities and material classes. *Journal of vision, 13*(8), 9-9.

Fleming, R. W. (2017). Material perception*. Annual review of vision science*, 3, 365-388.

Fujisaki, W., Tokita, M., & Kariya, K. (2015). Perception of the material properties of wood based on vision, audition, and touch. *Vision research*, *109*, 185-200.

Haghiri, S., Rubisch, P., Geirhos, R., Wichmann, F., & von Luxburg, U. (2019). Comparison-based framework for psychophysics: Lab versus crowdsourcing. *arXiv preprint arXiv:1905.07234*.

Haghiri, S., Wichmann, F. A., & von Luxburg, U. (2020). Estimation of perceptual scales using ordinal embedding. *Journal of vision*, *20*(9), 14-14.

Haindl, M., & Filip J. (2013). Visual Texture: Accurate Material Appearance Measurement, Representation and Modeling*. Advances in Computer Vision and Pattern Recognition, Springer-Verlag London*

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour*, *4*(11), 1173-1185.

Josephs, E. L., Hebart, M. N., & Konkle, T. (2023). Dimensions underlying human understanding of the reachable world. *Cognition*, *234*, 105368.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, *15*(2), 155-163.

Künstle, D. E., von Luxburg, U., & Wichmann, F. A. (2022). Estimating the perceived dimension of psychophysical stimuli using triplet accuracy and hypothesis testing. *Journal of Vision*, *22*(13), 5-5.

Lewin, M., & Goldstein, I.S. (1991). Wood Structure and Composition, *International Fiber Science and Technology, CRC Press*

Manuel, A., Leonhart, R., Broman, O., & Becker, G. (2015). Consumers' perceptions and preference profiles for wood surfaces tested with pairwise comparison in Germany. *Annals of forest science*, *72*(6), 741-751.

Marlow, P. J., Kim, J., & Anderson, B. L. (2012). The perception and misperception of specular surface reflectance*. Current Biology, 22*(20), 1909-1913.

Maskey, M., & Newman, T. S. (2021). On measuring and employing texture directionality for image classification. *Pattern Analysis and Applications*, *24*(4), 1649-1665.

McCamy, C. S. (1996). Observation and measurement of the appearance of metallic materials. Part I. Macro appearance. *Color Research & Application*, *21*(4), 292-304.

Motoyoshi, I., Nishida, S. Y., Sharan, L., & Adelson, E. H. (2007). Image statistics and the perception of surface qualities. *Nature, 447*(7141), 206-209.

Muttenthaler, L., Zheng, C. Y., McClure, P., Vandermeulen, R. A., Hebart, M. N., & Pereira, F. (2022). VICE: Variational Interpretable Concept Embeddings. *Advances in Neural Information Processing Systems*, *35*, 33661-33675.

Nakamura, M., Masuda, M., & Shinohara, K. (1999). Multiresolutional image analysis of wood and other materials. *Journal of wood science*, *45*, 10-18.

Nicodemus, F.E., Richmond, J.C., Hsia, J.J., Ginsburg, I.W., & Limperis, T. (1977). Geometrical considerations and nomenclature for reflectance. *NBS Monograph 160*, pp. 1-52

Nordvik, E., Schütte, S., & Broman, N. O. (2009). People's perceptions of the visual appearance of wood flooring: A kansei engineering approach. *Forest products journal*, *59*(11-12), 67-74.

Paulun, V. C., Schmidt, F., van Assen, J. J. R., & Fleming, R. W. (2017). Shape, motion, and optical cues to stiffness of elastic objects. *Journal of vision, 17*(1), 20-20.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, *40*, 49-70.

Rao, A. R., & Lohse, G. L. (1996). Towards a texture naming system: Identifying relevant dimensions of texture. *Vision Research*, *36*(11), 1649-1669.

Schmidt, F., Hebart, M. N., & Fleming, R. W. (2022). Core dimensions of human material perception. PsyArXiv. doi:10.31234/osf.io/jz8ks

Sharan, L., Liu, C., Rosenholtz, R., & Adelson, E. H. (2013). Recognizing materials using perceptually inspired features. *International journal of computer vision, 103*, 348-371.

Sharan, L., Rosenholtz, R., & Adelson, E. (2009). Material perception: What can you see in a brief glance?. *Journal of Vision, 9*(8), 784-784.

Sharan, L., Rosenholtz, R., & Adelson, E. H. (2014). Accuracy and speed of material categorization in real-world images. *Journal of vision, 14*(9), 12-12.

Tamura, H., Mori, S., & Yamawaki, T. (1978). Textural features corresponding to visual perception. *IEEE Transactions on Systems, man, and cybernetics*, *8*(6), 460-473.

Tanaka, M., & Horiuchi, T. (2015). Investigating perceptual qualities of static surface appearance using real materials and displayed images. *Vision research*, *115*, 246-258.

Van Assen, J. J. R., Barla, P., & Fleming, R. W. (2018). Visual features in the perception of liquids. *Current biology, 28*(3), 452-458.

Wan, Q., Li, X., Zhang, Y., Song, S., & Ke, Q. (2021). Visual perception of different wood surfaces: an event-related potentials study. *Annals of Forest Science*, *78*, 1-18.

Wendt, G., Faul, F., & Mausfeld, R. (2008). Highlight disparity contributes to the authenticity and strength of perceived glossiness. *Journal of Vision, 8*(1), 14-14.

Wendt, G., Faul, F., Ekroll, V., & Mausfeld, R. (2010). Disparity, motion, and color information improve gloss constancy performance. *Journal of vision, 10*(9), 7-7.

Wiebel, C. B., Valsecchi, M., & Gegenfurtner, K. R. (2013). The speed and accuracy of material recognition in natural images. *Attention, Perception, & Psychophysics, 75*, 954-966.

Supplementary material

1. VICE algorithm training

We tested the VICE model on 76 different combinations of input parameters such as et, spike, slab, pi and distribution (gaussian, laplace) (cf. Muttenthaler, Zheng, McClure, Vandermeulen, Hebart, & Pereira, 2022). Results are shown in Fig. S1(a), where the tested models are rank ordered according to test accuracy (red), with the corresponding training accuracy (blue). The converged models are highlighted as circles. Fig. S1(b) shows that the number of dimensions is relatively stable, within a range between 5 to 14 and a typical value of 10 dimensions.

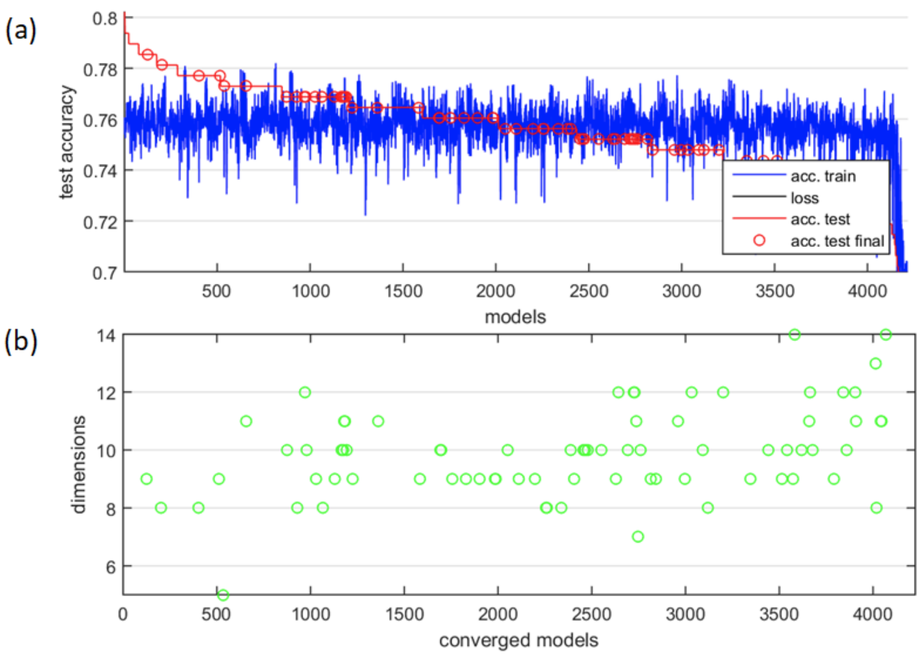


Fig. S1 Results of grid search across different VICE model parameters. (a) Model accuracies on train (blue) and test (red) sets (across all tested models) sorted according to accuracy on test set (red), and (b) corresponding obtained numbers of dimensions for converged models (also denoted as circles in (a)).

Fig. S2 shows the training process of the best performing converged model with the highest accuracy.

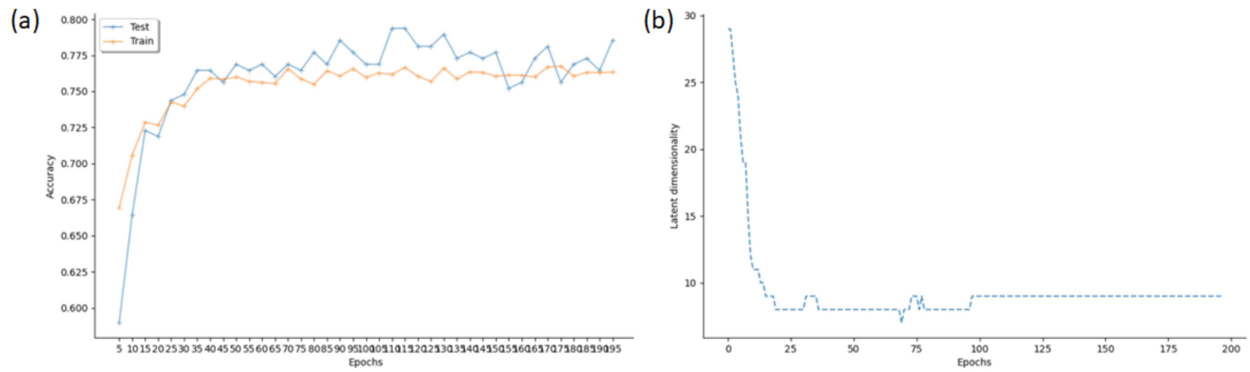


Fig. S2 Training process of the best performing model. (a) Model accuracy on the training (blue) and test (orange) dataset, (b) dimensionality reduction over 200 epochs of VICE algorithm.

2. Louvain community detection

We also applied the community detection method (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008) on the estimated similarity matrix. The resulting three clusters visualised in Fig. S3 can be interpreted as (1) contrast/roughness, (2) non-directional/low frequency, (3) directional/high frequency modes. These results are in agreement with the results of hierarchical clustering and MDS analysis.

Fig. S3 Clustering based on the Louvain community detection method divided the material samples into three clusters. See supplementary video [movie_Louvain.avi].

3. Similarity judgement data dimensionality analysis

As the dimensionality of our dataset is unknown, we follow a recent approach by (Künstle, von Luxburg, & Wichmann, 2022) to estimate the number of perceived dimensions from triplet experiments, based on triplet embedding accuracy. When splitting our triplet dataset into 90% training and 10% test samples, we obtain an ordinal Euclidean embedding (Haghiri, Wichmann, & von Luxburg, 2020) for the perceptual ratings. This procedure always leads to a decreasing triplet error (cross-validated on the validation set) with an increasing number of dimensions until a sufficient number of dimensions has been reached. We ran a cross-validation with 10 repetitions, resulting in a drop of accuracy with more than 6 dimensions. This suggests that inherent dimensionality of our dataset is close to 6 perceptual dimensions. Note that our analysis shown in Fig. S1(b) reports a dimensionality of the typical estimated similarity embedding between 8 and 10 dimensions. This seems to contradict the estimate of the inherent dimensionality of 6 as reported above (and shown in Fig. S4). However, our linear regression analysis (blue bars in Fig. 12(a)) suggests that several of our similarity dimensions (namely dimension 7) cannot be reliably predicted from the appearance ratings, which might suggest that: (1) our rating dimensions do lack some important visual features, or (2) the number of representational dimensions is lower than the estimate of the VICE algorithm. In favour of the latter, the factor loadings of individual dimensions (Fig. 3(b)) show a drop in loadings for dimensions higher than 5. Also, when using PCA on the rating data to test whether intercorrelations (Fig. 8(b)) allow us to reduce the dimensionality, we end up with not more than 6 dimensions.

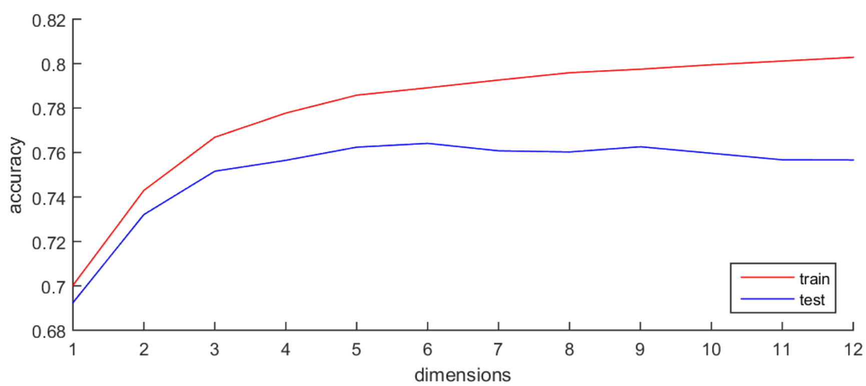


Fig. S4 Triplet ordinal embedding error as a function of the number of dimensions for training and test set of triplets from our similarity experiment.

4. Rating experiment details

The interface of the rating experiment is shown in Fig. S5.

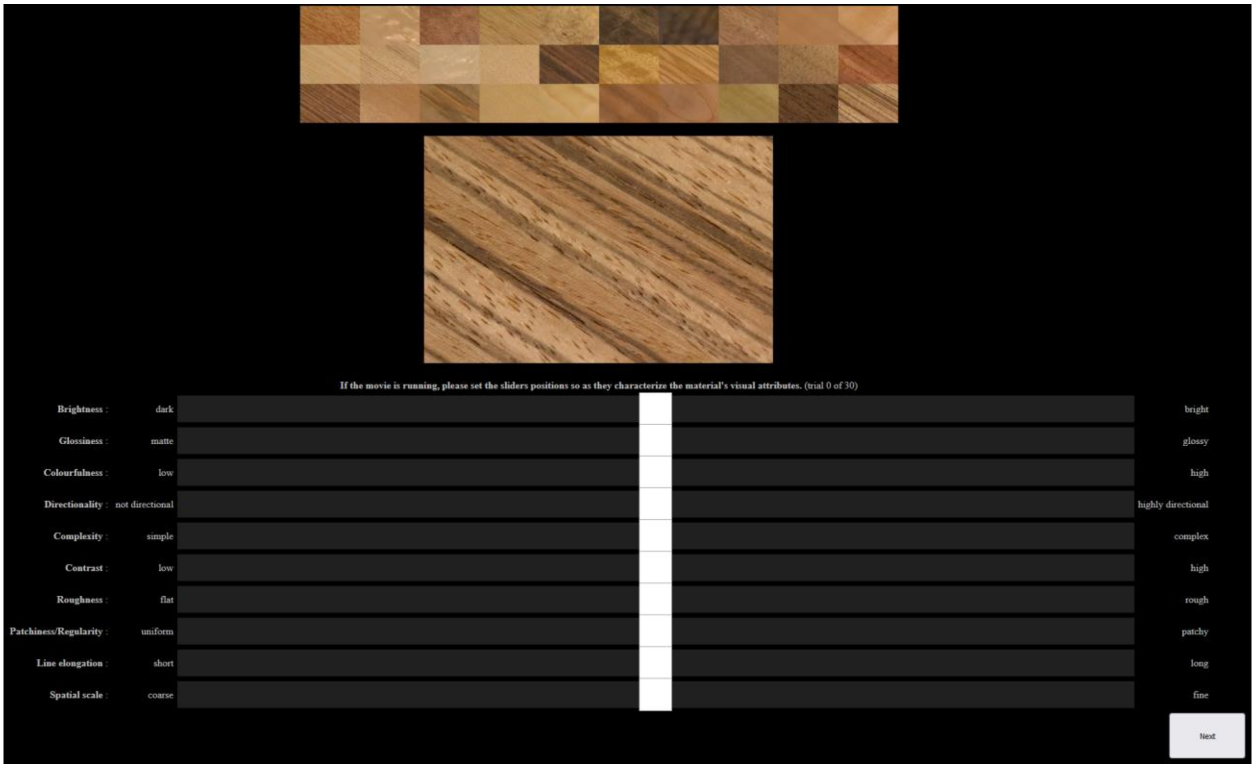


Fig. S5 Example stimulus frame from the rating experiment.

Instructions of the rating experiment were as follows: "*Below the video, there are 10 sliders for the visual material attributes. Your task is to adjust the slider position for each material. You may consider the appearance of the other materials (at the top of the screen) to choose your rating appropriately within the range we are testing*."

The following visual attributes are evaluated:

1. **Brightness** - how bright is the material?
2. **Glossiness** - how shiny is the material?
3. **Colourfulness** - how colourful is the material?
4. **Directionality** - presence of directional structures in the texture
5. **Complexity** - how complex are the patterns on the surface?
6. **Contrast** - difference in brightness of surfaces patterns
7. **Roughness** - smoothness of surface profile, range of surface heights
8. **Patchiness/Regularity** - how uniform is the pattern?
9. **Line elongation** - are line elements shorter dashes or extended lines?
10. **Spatial scale** - are patterns large and broad, or small and fine?

5. Rating results analysis

Mean values of participant responses for each resting attribute, and normalised distribution of allpooled ratings is shown in Fig. S6.

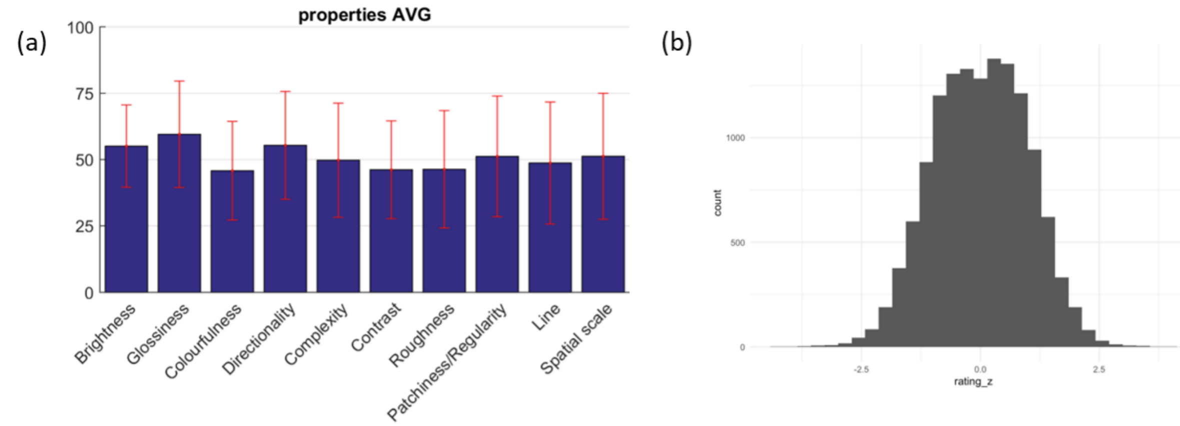


Fig. S6 Rating data analysis. (a) Mean values of participant responses across all materials with SD values, and (b) normalised distribution of all pooled ratings.

To evaluate the consistency between participants, we correlate the ratings of each participant within a scale to the corresponding mean rating (Fig. S7). Although overall correlations are pretty high, there is a heavy tail towards zero and even some negative correlations. Also, the consistency between participants varies between rating dimensions, for example, with more consistent judgements for brightness (stronger correlations and less variability). Tab. S1 shows intra-class correlations for individual rating dimensions.

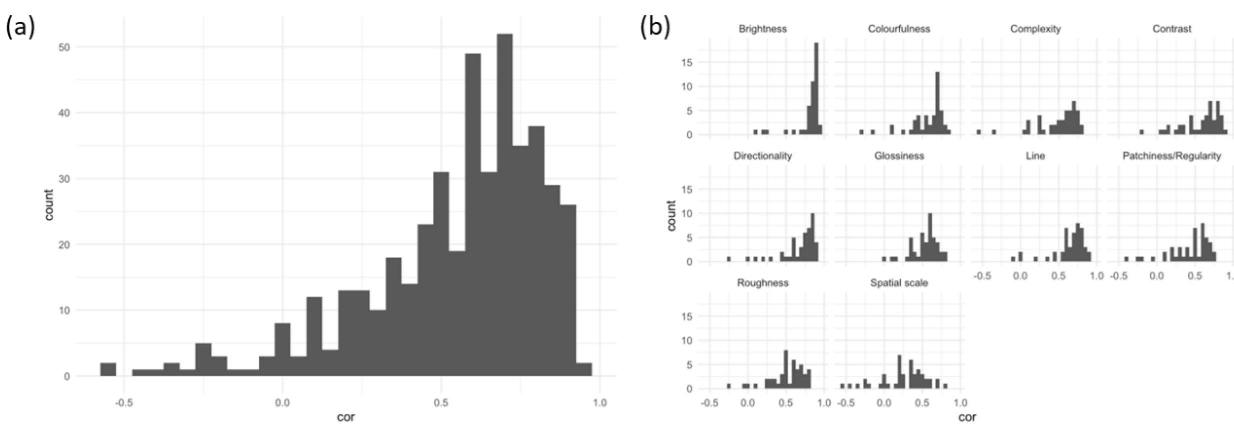


Fig. S7 Correlations between individual ratings of participants to the mean ratings, within each rating dimension. (a) Results across all attributes and (b) within individual dimensions.

Tab.S1 Intra-class correlations for individual rating dimensions.

rating dimension	**single random raters**	**average random raters**
brightness	0.618	0.986
glossiness	0.219	0.927
colourfulness	0.262	0.941
directionality	0.416	0.970
complexity	0.197	0.917
contrast	0.301	0.951
roughness	0.220	0.927
patchiness/regularity	0.164	0.898
line	0.386	0.966
spatial scale	0.041	0.659

6. Samples alignment along MDS dimensions

The MDS analysis distributed our 30 samples to three dimensional space. Distribution of samples along these dimensions is shown in Fig. S8, where red points represent MDS of VICE similarity model and blue MDS of rating attributes (the first two rows) and computational statistics (the third row).

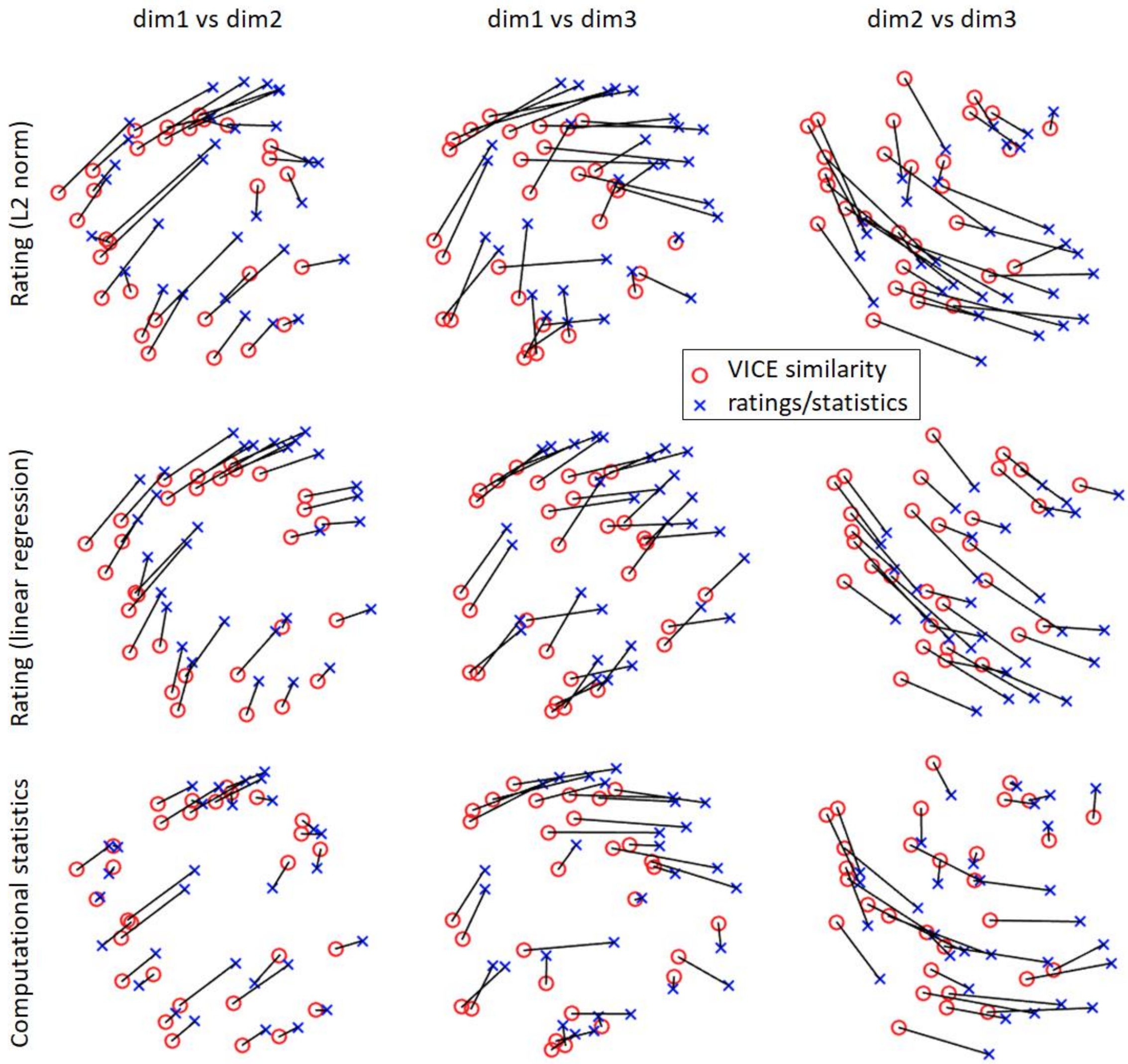


Fig. S8 Procrustes alignment of MDS dimensions computed from similarity matrices. (red) VICE model similarity MDS, (blue) MDS of similarity matrix obtained by L2- norm of rating attributes (the first row), linear regression of rating attributes similarity matrices (the second row), and computational statistics MDS (the third row).

List of supplementary movies

1. [movie_samples_stimuli.avi] - 30 test wood video sequences used in the experiments
2. [movie_similarity_vs_rating.avi] - rank ordered samples (left) according to loadings values of similarity dimensions, (right) mean rating attributes (the five closes and 5 the most distant)
3. [movie_MDS_simmat_linreg.avi] - distribution of samples along three MDS dimensions (top) for similarity judgements, (bottom) for rating study
4. [movie_MDS_simmat_stat.avi] - distribution of samples along three MDS dimensions (top) for similarity judgements, (bottom) for computational statistics obtained from image sequence.
5. [movie_similarity_scaled.avi] - rank ordered samples scaled according to loadings values of similarity dimensions
6. [movie_Louvain.avi] - result of community detection using Louvain method (computed from similarity matrices), distributing samples to three clusters for (top) similarity judgements and (bottom) rating study.