Contribution of Low, Mid and High-Level Image Features of Indoor Scenes in Predicting Human Similarity Judgements

Anastasiia Mikhailova[1(✉)], José Santos-Victor[1], and Moreno I. Coco[2]

[1] Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal
anastasiia.mikhailova@tecnico.ulisboa.pt
[2] Sapienza, University of Rome, Rome, Italy

Abstract. Human judgments can still be considered the gold standard in the assessment of image similarity, but they are too expensive and time-consuming to acquire. Even though most existing computational models make almost exclusive use of low-level information to evaluate the similarity between images, human similarity judgements are known to rely on both high-level semantic and low-level visual image information. The current study aims to evaluate the impact of different types of image features on predicting human similarity judgements. We investigated how low-level (colour differences), mid-level (spatial envelope) and high-level (distributional semantics) information predict within-category human judgements of 400 indoor scenes across 4 categories in a Four-Alternative Forced Choice task in which participants had to select the most distinctive scene among four scenes presented on the screen. Linear regression analysis showed that low-level ($t = 4.14$, $p < 0.001$), mid-level ($t = 3.22$, $p < 0.01$) and high-level ($t = 2.07$, $p < 0.04$) scene information significantly predicted the probability of a scene to be selected. Additionally, the SVM model that incorporates low-mid-high level properties had 56% accuracy in predicting human similarity judgments. Our results point out: 1) the importance of including mid and high-level image properties into computational models of similarity to better characterise the cognitive mechanisms underlying human judgements, and 2) the necessity of further research in understanding how human similarity judgements are done as there is a sizeable variability in our data that it is not accounted for by the metrics we investigated.

Keywords: Image similarity · Scene semantics · Spatial envelope · SVM · Hierarchical regression

This research was supported by Fundação para a Ciência e Tecnologia with a PhD scholarship to AM [SFRH/BD/144453/2019] and Grant [PTDC/PSI-ESP/30958/2017] to MIC.

© Springer Nature Switzerland AG 2022
A. J. Pinho et al. (Eds.): IbPRIA 2022, LNCS 13256, pp. 505–514, 2022.
https://doi.org/10.1007/978-3-031-04881-4_40

1 Introduction

Evaluating the similarity of visual information is an important challenge for computer vision that finds its application in object recognition [1], template matching [2], generalization of robot's movement in space [3,4], reverse search of products or images [5] and many other practical applications.

As image similarity is a difficult computer vision task, humans are still often required to provide their assessment as they can successfully evaluate image similarity even in noisy conditions [6]. But relying on humans is expensive and time-consuming because the number of judgements needed grows quadratically according to the number of evaluated pair of images. However, even though the nature of human judgements is subjective, it may still be possible to frame this problem computationally by considering different types of metrics that could characterise an image. We could identify three types of features on which human similarity judgements may be based: low-level visual information (e.g., pixels, colour), mid-level structural information that is specific to scenes (spatial envelope) and high-level semantic information (e.g., object and scene concepts) [7].

At present, the majority of computational models assessing image similarity rely only on low-level visual image properties (e.g., [1,2]) as it is the easiest information that could be extracted from an image. Thus, other types of information, as those we listed just above, are often not accounted for when calculating their similarity. Hence, to develop more reliable computational models of image similarity, it is important to understand the interplay of other types of mid and high-level information when humans are asked to perform similarity judgements.

Of the existing literature we are aware of, only a few studies are attempting to model the features that drive similarity judgements in humans. The study by [8] is one such example, where authors trained Sparse Positive Similarity Embedding (SPoSE) [9] on more than 1.5 million similarity judgements on images of objects in an odd-one-out task, where participants needed to exclude the most distinct image out of three. Their results, which are based on interpreting the SPoSE dimensions, highlighted that humans rely on both low and high-level features of images to provide their similarity judgments.

The study by Hebart and colleagues' [8], which we briefly touched upon, was done on individual objects and measured their similarities across different semantic categories (e.g., toy vs kettle). Even though it is common to operationalise this task between categories, it is important to frame it also in the within-category context [10] and expand it to the larger scope of naturalistic scenes. A within-category similarity task is, in fact, harder than a between-category and probably require access and use of a variety of features to be accurately performed. Moreover, by looking at naturalistic scenes instead of individual objects, we can also evaluate the role that mid-level features, such as their spatial envelope [11], play in similarity judgements.

In the current study, we precisely aimed to investigate how humans perform a within-category similarity judgement of naturalistic scenes and evaluated the importance of their low, mid and high-level features in such a task. Our goal is to provide a better understanding of the cognitive factors implied in human

similarity judgments while evaluating the predictivity of different computational metrics that may be implicated in performing this task.

2 Study of Similarity Judgements

2.1 Image Dataset

We used the ADE20K [12] and SUN [13] databases and selected images following these criteria:

(1) contained full annotation of object labels, which is necessary to compute their high-level semantic similarity (see next section for details on this metric).
(2) did not contain animate objects (e.g., people or animals) as it makes them more distinctive [14].
(3) at least 700 pixels in width or height to ensure a reasonable resolution quality.

According to these criteria, we kept 4 categories with more than 200 images each: bathroom, bedroom, kitchen and living room (1,319 images). These images were further filtered down to 100 images per category for the study to fit within a reasonable time frame (i.e., under one hour), and selected by excluding the most dissimilar images based on low, mid, and high-level measures described below. Thus, the final dataset comprised 400 images in 4 different semantic categories.

2.2 Low, Mid, and High-Level Image Features

As mentioned above, we can identify three possible features on which similarity of images can be estimated: low (e.g., colour) mid (e.g., spatial envelope) and high-level (e.g., semantics).

A very simple measure of low-level similarity between images can be obtained from their colour at the pixel level [15] (Eq. 1). Specifically, images can be represented as three-dimensional matrices in RGB space and obtain a pairwise similarity by simply subtracting matrices. Then, the sum of all differences at the pixel level, raised to the power of two, can be taken as a single pairwise similarity value (sum of square differences, SSD, or l^2-norm) measure.

$$S_{SSD} = \sum (I[n, m, k] - J[n, m, k])^2 \tag{1}$$

where I and J indicate the indices of two images being compared, $[n, m, k]$ points to a pixel position in three dimensional matrix of RBG image and S_{SSD} is the similarity distance of the image I and J according to the SSD measure.

Mid-level information of an image can be obtained from the spatial distribution of low-level spectral information in it. We follow the classic study by [11] and compute the power spectral density of each image divided into 8×8 regions in 8 different orientations. Thus, each image is represented as a vector of 512

values (64 regions × 8 orientations), a spatial envelope, which is also known to convey coarse information about its semantic category [16]. Then, the similarity between images is computed in pairwise fashion as the sum of the square difference of the GIST vectors:

$$S_{GIST} = \sum (GIST_i - GIST_j)^2 \tag{2}$$

where GIST is a vector of 512 values of GIST descriptor, i and j are indices of two images being compared and S_{GIST} is the similarity distance of the image i and j according to the GIST measure.

Finally, high-level semantic information of a scene can be approached by considering the objects therein as its conceptual building blocks and their co-occurrence statistics as the metric. This approach assumes that the more objects scenes share, the more semantically similar they will be [17] and we utilised the method conceived by Pennington and colleagues [18] (Global Vectors model, GloVe) to operationalise this measure. Specifically, the labels of objects in each image were used to create a co-occurrence matrix of objects across all images in each of four categories. This matrix was transformed using a weighted least squares regression into a vector representation of each object (50 dimensions), and each image was represented as the sum of vectors of objects contained therein. Then, the pairwise similarity between images was calculated as a reverse cosine distance of the GloVe vectors:

$$S_{GloVe} = 1 - cos(GloVe_i, GloVe_j) \tag{3}$$

where GloVe is a vector of 50 values of GloVe descriptor, i and j are indices of two images being compared and S_{GloVe} is the similarity distance of the image i and j according to the GloVe measure.

These three measures of similarity described above were applied to all images belonging to a certain semantic category and the value for each image represented as its distance to all other images in the given category. The three measures were then normalized using a z-score transformation to make them more directly comparable.

Additionally, we considered average feature maps from VGG-16 neural network trained on Places365 image set [19] from the first and the last convolutional layers as the proxy for low- and high-level feature information of the scenes. The average maps were compared using Kullback-Leibler divergence between activation maps of the images but as it did not lead to significant results, we would not report this data here.

2.3 Similarity Task

Forty participants viewed the stream of images presented four at a time, all from the same semantic category, and asked to click on the one that was, in their opinion, the most different in the set, so performing a Four-Alternative Forced Choice task (see Fig. 2 for the examples of design and human judgements

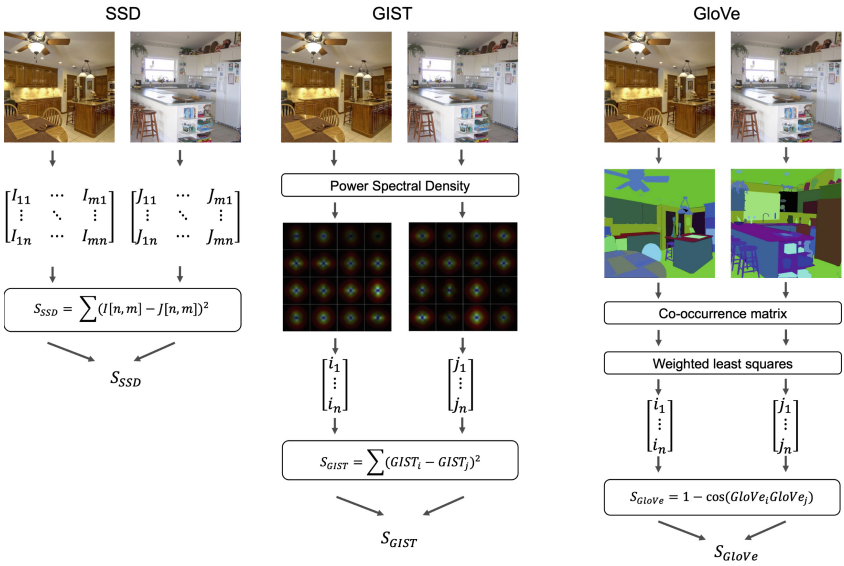

Fig. 1. Illustration of the similarity measures: SSD, GIST and GloVe.

output). For each trial/image combination we collected 10 subjective similarity judgements. To test the predictivity of GIST or GloVe in explaining similarity judgments, the four images in each trial were arranged such that there was one more distinctive (target) selected on the basis of its maximum distance to the other three (foils) images according to GIST or GloVe measures of similarity. This resulted in two different experimental conditions based on either GIST or GloVe, respectively. Additionally, we ensured that all images were properly counterbalanced in their possible combinations by creating two further lists from each condition.

The participants had a time limit of 6 s to provide their judgment, otherwise, a null response was logged in, and the trial was excluded from the analysis. Out of 4,000 trials (25 trials per category $\times$ 4 categories $\times$ 40 participants), we excluded 14.5% of the trials because of timeout. The Gorilla platform [20] was used to implement the study and collect the data.

2.4 Analysis

The data of this study were analysed in two ways.

The first analysis used linear-mixed effect modelling (using the R package lme4 [21]) to investigate how and to what extent the probability of excluding a target image can be predicted by similarities in their SSD, GIST and GloVe with the foils. The probability of an image to be excluded was computed by averaging similarity judgements across all participants for the trial in which such an image appeared. So, it was defined as the proportion of participants selecting a certain

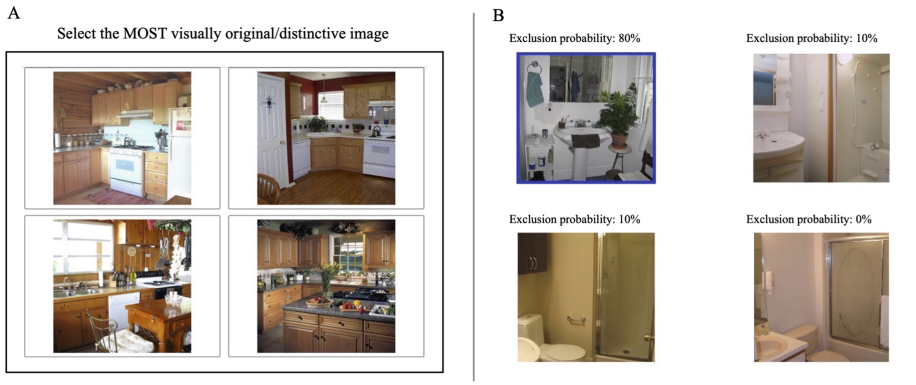

Fig. 2. A. Example of the Four-Alternative Forced Choice task to acquire human similarity judgements. B. Illustration of exclusion probability based on human similarity judgments within a trial.

image among three other images as the most distinctive one at each trial. Thus, a high exclusion probability indicated that the participant found this image to be the most distinctive relative to the other images (see Fig. 2). This measure, the exclusion probability of an image, was used as the dependent variable for the regression analysis as it reflects the agreement or consistency between the participants in their similarity judgment.

In the second analysis, we used Support Vector Machine (SVM) classifiers to examine the predictivity of the similarity metrics in approximating human judgments (binary classification). For this analysis, exclusion probability was expressed as a binary variable whereby a 1 indicated the most selected scene in each trial (i.e., an exclusion probability above 0.25, which is chance level) and 0 otherwise. Four different SVM models were created to predict the exclusion probability using each metric as an independent predictor (GIST, GloVe or SSD), and one more SVM was built as an additive (linear) combination of all predictors (GIST, GloVe and SSD) to explore the joint contribution of low, mid and high level information.

SVMs were trained and tested with a 70/30% ratio on 10-fold cross-validation over 100 iterations. We used the *ksvm* function with default settings from kernlab R package [22] to implement the SVM. The *predict* and *confusionMatrix* functions were used to extract the accuracy, f-score, and other measures that characterised the performance of the SVM models.

We compared the performance of the SVM (all predictors combined: GIST, GloVe and SSD) against the human data by averaging all SVM predictions (i.e., 100 iterations * 10 folds) for each image, and so obtain a dependent measure more directly comparable to the human judgments, and used independent-sample t-test as well as the Kolmogorov-Smirnov test to assess whether differences between human judgments and model predictions were significant.

2.5 Results

To provide a descriptive analysis of the task and measure the consistency in human judgements, we calculated the percentage of inter-observer agreement. On a trial-by-trial basis, we calculated the percentage of observers to pick the most selected image in the trial. We observed a 49% inter-observer agreement, which is much above change (i.e., 25% in this task) but still shows a certain degree of variability among them.

When looking at the linear mixed regression analysis, we found that all three measures GIST, GloVe and SSD had a significant effect on the probability of exclusion (see Table 1), which indicates that humans rely on all three types of features to perform similarity judgements.

Table 1. Results of linear regression analysis, where the effect of GIST, SSD and GloVe on probability of image exclusion in Four-Alternative Forced Choice task is tested.

Predictor	Beta coefficient	t-value	SE	p-value
GIST	0.11	4.14	0.01	$\mathbf{<0.001^{***}}$
SSD	0.08	3.22	0.01	$\mathbf{<0.01^{***}}$
GloVe	0.05	2.07	0.01	$\mathbf{<0.04^{*}}$

Turning onto the SVM predictions, we note that the model combining all 3 similarity measures predicts human similarity judgements best (i.e., 56%) and significantly better than models with single predictors (refer to Table 2).

Table 2. Accuracy of SVM models averaged across 10 fold and 100 random initialisation and the t-test comparison of single predictor models relative to the full model (GIST, GloVe and SSD).

Model	Accuracy	t-score	SE	p-value
GIST + GloVe + SSD	55.9%			
SSD	54.9%	−3.08	0.32	$\mathbf{<0.01^{**}}$
GIST	54.2%	−5.15	0.32	$\mathbf{<0.001^{***}}$
GloVe	53.9%	−6.17	0.32	$\mathbf{<0.001^{***}}$

When comparing the performance of SVM and humans, we observe no significant difference under the t-test ($t = 0.64$, $p = 0.52$), which instead is significant under the Kolmogorov-Smirnov test ($D = 0.35$, $p < 0.001$) indicating that the distributions of exclusion probability according to humans and SVM are significantly different even though, they have a similar mean.

3 Discussion

Current research shows that humans rely on both low- and high-level image information to judge their similarities [8], even though most computational models seem to be largely based only on low-level image properties [10]. Moreover, most of this research focused on similarity between categories rather than within, which is a much harder task to solve (e.g., assessing the similarities between different kitchens) [10]. The goal of the present study was precisely to evaluate the role played by low, mid and high-level image features on driving within-category human similarity judgements. Our aims were two-fold: expand our understanding of the cognitive processes humans rely on when assessing the similarity of images and evaluate the computational predictivity of low, mid and high-level metrics in approximating human performance.

Our data shows that all types of image features (low, mid and high) contribute to explaining how humans perform similarity judgements. These results corroborate findings by Hebart and colleagues' [8] but in naturalistic scenes, and especially in within-category context, which is not typically done in this research field. To provide a further, mainly qualitative, comparison with this study, we retrained their SPoSE model on our within-category data and found that the embedding dimensions obtained from the model do not seem to have a significant predictive effect on similarity judgements ($t = 1.53$, $p = 0.13$). Such results tentatively indicate that even the most recent computational models of image similarity are not well-suited yet to assess more fine-grained within-category similarities. We acknowledge, however, that our results obtained using SPoSE might be limited due to the significantly smaller sample size and call for further research on within-category similarity.

Another theoretical contribution of the current study was to show the importance of mid-level features in this task (i.e., GIST), which are often neglected by previous research. This finding confirms the necessity to incorporate other types of features into computational models of image similarity and so aiming to better approximate the way humans may be solving this task [6]. When looking at the best predictivity of human performance using SVM, we found that all features should be used to achieve the highest performance (i.e., 56%). We note, however, that such accuracy remains still pretty low and so there may likely be other factors, and perhaps also more efficient metrics, that should be used to better model the performance of humans in this task. Crucially, the accuracy of our SVM models remains the same even when using other measures of low- and high-level information. In an exploratory analysis not presented at length in the current study, we also used the average activation map from the first layer of the VGG-16 neural network [19] as a proxy to low-level features and of the last layer as a proxy to high-level features. The pairwise similarity between images, in this case, was computed as Kullback-Leibler divergence between activation maps. SVM model trained on these features still produces a prediction performance of 54%. We acknowledge that there are more modern sophisticated deep-neural network models that could be used to derive metrics of image similarity and so better understand the nature of human judgement, compared to the rather

simple metrics this work utilise. This consideration reinforces the call to further research into other metrics to compute image similarity, and perhaps aspects in the task, that may better uncover the cognitive underpinnings of humans similarity judgement. Finally, we also note that the inter-observer agreement is also relatively low (49%) so indicating quite some disagreement between human judgements. We believe that precisely understanding root causes of their disagreements would indirectly help us capturing the processes that would make them agree on their judgements instead.

Finally, we realise that the scope of the current study is restricted to a within-category similarity scenario and considers only few indoor categories; and agree that to evaluate a wider spectrum of low-, mid- and high level differences on similarity judgements outdoor scenes should be considered.

4 Conclusion

First, our results demonstrate the importance of incorporating mid-level spatial and high-level semantic image information along commonly used low-level visual information when modelling similarity judgements. Secondly, we bring attention to the investigation of within-category similarity judgements, as it is an intrinsically harder task for similarity models to solve. Lastly, we point out that more research is needed to understand all aspects of human similarity judgments as our modelling, which incorporated low-mid-high level information, is still not sufficient to account for the overall variability observed across human judgements.

References

1. Sampat, M.P., Wang, Z., Gupta, S., Bovik, A.C., Markey, M.K.: Complex wavelet structural similarity: a new image similarity index. IEEE Trans. Image Process. **18**(11), 2385–2401 (2009)
2. Zhang, Y., Zhang, C., Akashi, T.: Multi-scale Template Matching with Scalable Diversity Similarity in an Unconstrained Environment (2019)
3. Wu, A., Piergiovanni, A.J., Ryoo, M.S.: Model-based behavioral cloning with future image similarity learning. In: Conference on Robot Learning, pp. 1062–1077 (2020)
4. Wang, L., et al.: Image-similarity-based convolutional neural network for robot visual relocalization. Sens. Mater. **32**, 1245–1259 (2020)
5. Bell, S., Bala, K.: Learning visual similarity for product design with convolutional neural networks. In: ACM Trans. Graph. (TOG) **34**(4), 1–10 (2015)
6. Silva, E.A., Panetta, K., Agaian, S.S.: Quantifying image similarity using measure of enhancement by entropy. In: Mobile Multimedia/Image Processing for Military and Security Applications 2007 6579, p. 65790U (2007)
7. Liu, Y., Gevers, T., Li, X.: Color constancy by combining low-mid-high level image cues. Comput. Vision Image Understanding **140**, 1–8 (2015)
8. Hebart, M.N., Zheng, C.Y., Pereira, F., Baker, C.I.: Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. Nat. Hum. Behav. **4**(11), 1173–1185 (2020)

9. Zheng, C.Y., Pereira, F., Baker, C.I., Hebart, M.N.: Revealing interpretable object representations from human behavior. In: International Conference on Learning Representations (2018)
10. Wang, J., et al.: Learning fine-grained image similarity with deep ranking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1386–1393 (2014)
11. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. Int. J. Comput. Vision **42**(3), 145–175 (2001)
12. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 633–641 (2017)
13. Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., Torralba, A.: Sun database: large-scale scene recognition from abbey to zoo. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3485–3492 (2010)
14. Bylinskii, Z., Isola, P., Bainbridge, C., Torralba, A., Oliva, A.: Intrinsic and extrinsic effects on image memorability. Vision Res. **116**, 165–178 (2015)
15. Ulysses, J. N., Conci, A.: Measuring similarity in medical registration. In: IWSSIP 17th International Conference on Systems, Signals and Image Processing (2010)
16. Oliva, A., Torralba, A.: Building the gist of a scene: the role of global image features in recognition. Progress Brain Res. **155**, 23–36 (2006)
17. Sadeghi, Z., McClelland, J.L., Hoffman, P.: You shall know an object by the company it keeps: an investigation of semantic representations derived from object co-occurrence in visual scenes. Neuropsychologia **76**, 52–61 (2015)
18. Pennington, J., Socher, R., Manning, C. D.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
19. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: arXiv preprint arXiv:1409.1556 (2014)
20. Anwyl-Irvine, A.L., Massonnié, J., Flitton, A., Kirkham, N., Evershed, J.K.: Gorilla in our midst: an online behavioral experiment builder. Behav. Res. Methods **52**(1), 388–407 (2019). https://doi.org/10.3758/s13428-019-01237-x
21. Bates, D., Mächler, M., Bolker, B., Walker, S.: Fitting linear mixed-effects models using lme4. J. Stat. Softw. **67**(1), 1–48 (2015)
22. Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A.: kernlab-an S4 package for kernel methods in R. J. Stat. Softw. **11**(9), 1–20 (2004)