Does Free-sorting Provide a Good Estimate of Visual Similarity

Alasdair D. F. Clarke
Institute for Language,
Cognition and Computation
University of Edinburgh
Informatics Forum
Edinburgh, UK
a.clarke@ed.ac.uk

Xinghui Dong
The Texture Lab
School of Mathematical and
Computer Sciences
Heriot-Watt University
Edinburgh, UK
xd25@hw.ac.uk

Mike J. Chantler
The Texture Lab
School of Mathematical and
Computer Sciences
Heriot-Watt University
Edinburgh, UK
m.j.chantler@hw.ac.uk

ABSTRACT

The majority of work on texture analysis in computer vision has concerned texture classification and segmentation, while the problem of measuring and modelling the visual similarity between pairs of textures has been relatively neglected. One likely reason for this is the difficulty in collecting subjective human similarity judgments over a large database of textures. A common approach is to carry out a free-sorting experiment to obtain a similarity matrix which can then be mapped onto a low dimensional space using techniques such as MDS or Isomap. This results in a Euclidean space in which textures are represented as points, and the distance between two points is taken to represent the perceptual visual dissimilarity between the associated pair of textures. However, it is unknown if such a metric can generalise to predict human texture judgements in other tasks, or even if similarity judgements are metric at all. In this study we investigate this question by carrying out an experiment using a pair-of-pairs paradigm and compare these results to the predictions made by a low dimensional model (d = 3) obtained from a free-sorting experiment and find that it agrees with the judgements made by participants.

1. INTRODUCTION

Texture classification and segmentation have been extensively researched over the last thirty years and while extremely successful algorithms have been developed to address classification problems, the challenging problem of measuring perceived inter-class texture similarity has received less attention. For example, we want to be able to measure the perceptual difference between the pairs of textures shown in Figure 1. Tamura et al [10] were some of the first to con- sider the problem of computational similarity and criticised earlier work for overlooking this aspect of texture analysis:

> *However, [the] features are not obviously visual. Sometimes even random selection of features may give satisfactory accuracy in a classification problem, especially if they are orthogonal by accident. . . . our challenge is to develop the textural features approximating visual perception."*

***Predicting Perceptions: The 3rd International Conference on Appearance**, 17-19 April, 2012, Edinburgh, UK. Conference Proceedings Publication ISBN: 978-1-4716-6869-2, Pages: 17-20 @2012 Authors & Predicting Perceptions. All Rights Reserved.*

We believe that this criticism is as relevant today as it was 30 years ago.

One of the difficulties of analysing the visual similarity be- tween two textures is that, unlike with classification, there is no objective ground truth: perceived similarity can vary along a continuum, from “identical” to “completely different.” How consistent are human observers? Can we reliably measure the difference between such pairs of textures using computational methods? Furthermore, collecting empirical similarity data is a challenging and time consuming task in its own right. Different people are likely to have different opinions on how similar one texture is to another and, as argued by Heaps & Handel [5], context plays an important role. There is also evidence that people show a preference for making unimodal decisions when faced with stimuli that differ along several different dimensions [2]. Ideally, we would like to collect data that will give us insight into the structure of perceptual texture space, (analogous to the $L^*u^*v^*$ colour space). Is a dimensional model appropriate? Does the space consist of a collection of subspaces that would be best considered separately (textiles, rocks, etc.)? Can we accurately characterise all natural textures using a low number of dimensions as Rao & Lohse [7] suggest?

In the current study we will carry out a new similarity experiment using the pair of pairs paradigm and the PerTex texture collection [4]. The advantage of this method is that it provides more precise data with which to assess computational texture features. The downside is that we can only obtain data for a tiny fraction of all possible combinations (there are $4^{334} \approx$ 500 million in total), and each individual trial only tests a model’s ability to give the correct ordinal ranking rather than its performance as a metric with an interval scale that correlates with human perception.

We will also develop a low-dimensional perceptual metric from the grouping data provided with [4] and test how well it can account for the results of the pair-of-pairs experiment. As with Long & Leow [6], we do not try to attach perceptual features to our model, and we agree with Amadasun & King’s suggestion that there is no reason to think that there is a single meaningful set of global texture features that apply to all surfaces [1]. Our aim with creating a perceptual metric is simply to use it as a yardstick with which to measure the performance of computational algorithms and to investigate the degree to which human judgements can be approximated with a Euclidean metric.

2. PERCEPTUAL EXPERIMENTS

The PerTex collection [4] comprises of 334 textures and an associated similarity matrix (obtained from a free-sorting experiment with 30 participants), which gives an estimate of the visual similarity between all textures. If we want to rigorously evaluate the performance of different computational similarity algorithms,then we need a large number of samples in order to represent a realistic variety of surface textures and the different levels of similarity between them. If we use a small dataset that covers a large range of visual appearances, such as the Brodatz set, then we end up with a very sparse similarity matrix with very few pairs of textures that have any degree of similarity between them. However, the downside of using a large dataset is that the number of pairwise comparisons rapidly increases. Previous studies comparing different experimental paradigms for collecting similarity judgments used much smaller sets (< 50) and concluded that while free-sorting is the least time intensive method, the data obtained is not as precise as with other data collection methods[8, 3].

2.1 Creating a Perceptual Model

In order to construct a perceptual metric from the PerTex similarity matrix, S_{ij} , we first transform it into a dissimilarity matrix : $d_p(I_i, I_i) = 1 - S(I_i, I_i)$. Hence $d_p(I_i, I_i) = 0$ for all images I_i, and $d_p(I_i, I_j) = 1$ if none of the participants grouped images I_i and I_j together. We then use Isomap[11] to explore the extent to which a low dimensional metric space can accommodate the perceptual similarity matrix. Isomap works by first creating a graph based on short distances between points, ignoring longer distances. For our purposes, we use all $d_p(i, j) < 1$ to create the graph, treating the similarity scores for texture pairs that were never grouped together as missing values. The distance from one point to another is then defined as the shortest-path distance on the graph. MDS [9] is then applied to this transformed set of proximities to find a low dimensional embedding.

A three dimensional metric space can accommodate over half the variance in the human data (see Figure 2). Due to the nature of the free-sorting experiment, the empirical data is highly skewed and quantised and we would not expect a high correlation with a low number of dimensions (over 80% of the entries in the dissimilarity matrix have the maximal value of $d_p(i, j) = 1$). In the following section we will evaluate this model on a new set of empirical similarity judgements obtaining from a new experiment using a different paradigm. If the extrapolated values obtained from applying Isomap to the free-sorting similarity matrix are a good fit with human perception then we would expect the model to agree with the results of this new experiment.

2.2 Pairs of Pairs Experiment

In order to test the reliability of the perceptual model created above we carried out a new, independent experiment. 1000 pairs of pairs $\{\{a, b\}, \{c, d\}\}$ were randomly selected from the set of 334 textures (the only criteria was that $a = b$ and $c = d$). In each trial four textures were shown on the screen and a total of 20 participants were asked to judge whether the two textures on the left were more similar to one another than the pair on the right.

Figure 3 shows how consistent the human participants were, and how well d_{iso} does in the pairs of pairs task. All participants responded the same way for 12% of trials and the mean agreement between participants was 70.6% (std. err. = 1.65%). The perceptual model performed well, with a mean agreement of 67.6% with three dimensions. Increasing the number of dimensions only offered marginal improvements (68.6% for $d = 6$). Figure 2 suggests that a higher number of dimensions are necessary to accurately represent the grouping data. $d = 3$ gives $r = 0.55$ while $d = 8$ results in $r = 0.76$. This highlights the effect different data collection methods (ordinal versus ratio data) can have on analysis.

3. CONCLUSION

Unlike previous studies on texture similarity, we have compared the results from two different experiment designs to investigate if it is valid to generalise from the results of free-sorting experiments. While we do not claim that human similarity judgements are metric in nature, we do find that an empirically derived metric offers a good approximation to human behaviour in a pair-of-pairs task.

4. REFERENCES

[1] M. Amadasun and R. King. Textural features corresponding to textural properties. IEEE Transactions on Systems, Man and Cybernetics, 19:1264–1274, 1989.

[2] F. Gregory Ashby, Sarah Queller, and Patrica M. Berretty. On the dominance of unidimensional rules in unsupervised categorization. Perception & Psychophysics, 61:1178–1199, 1999.

[3] Tammo H. A. Bijmolt and Michel Wedel. The effects of alternative methods of collecting similarity data for multidimensional scaling. International Journal of Research in Marketing, 12:363–371, 1995.

[4] A. D. F. Clarke, F. Halley, A. Newell, L. D. Griffin, and M. J. Chantler. Perceptual similarity: a new texture challenge. In BMVC2011, 2011.

[5] C. Heaps and S. Handel. Similarity and features of natural textures. Journal of Experimental Psycholog.: Human Perception and Performance, 25:299–320, 1999.

[6] H. Long and W. K. Leow. Perceptual texture space improves perceptual consistency of computational features. In IJCAI'01 Proceedings of the 17th international joint conference on Artificial intelligence, 2001.

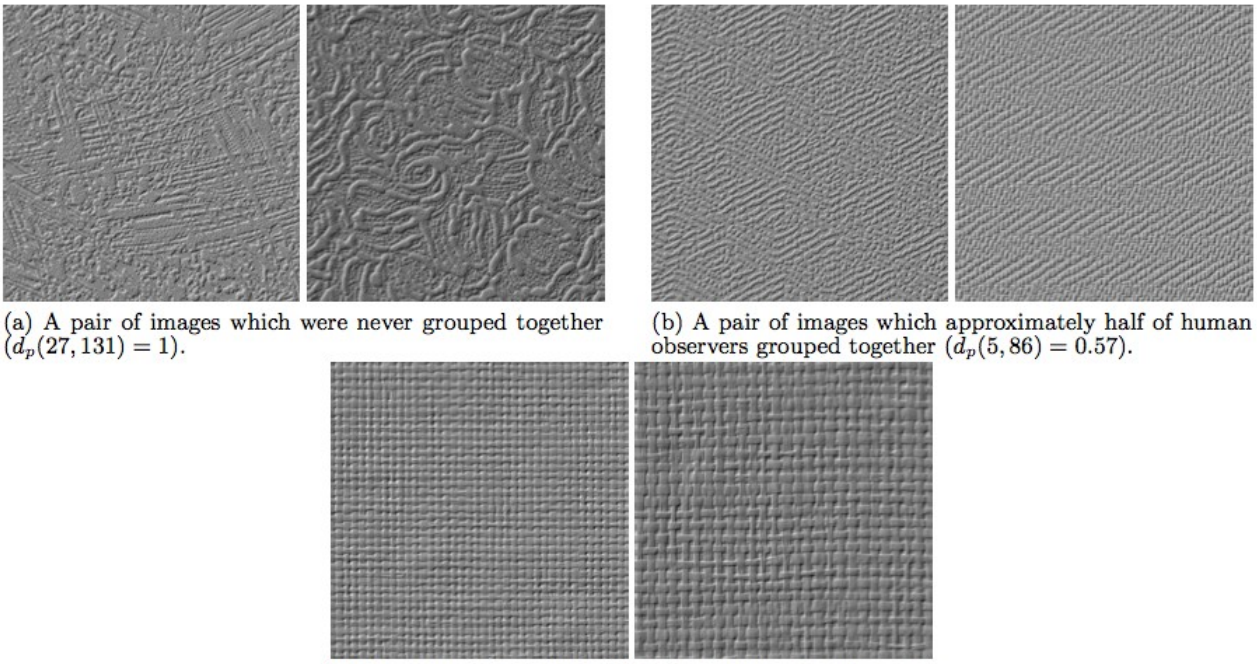
[7] A. R. Rao and G. L. Lohse. Identifying high level features of texture perception. CVGIP: Graph. Models Image Process., 55:218–233, May 1993.

[8] Vithala R. Rao and Ralph Katz. Alternative multidimensional scaling methods for large stimulus sets. Journal of Marketing Research, VIII:488–494, 1971.

[9] R. N. Shepard. Analysis of proximities: Multidimensional scaling with an unknown distance function. i. Psychometrika, 27:125–140, 1962.

[10] H. Tamura, S. Mori, and T. Yamawaki. Textural features corresponding to visual perception. IEEE Transactions on Systems, Man and Cybernetics, 8:460–473, 1978.

[11] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. Science, 290:2319–2323, 2000.



(a) A pair of images which were never grouped together ($d_p(27, 131) = 1$).

(b) A pair of images which approximately half of human observers grouped together ($d_p(5, 86) = 0.57$).

(c) A pair of images which all but one of human observers grouped together ($d_p(168, 176) = 0.07$).

Figure 1: Some examples of textures from the PerTex dataset with their pairwise similarity. Unlike the binary classification and segmentation tasks, the degree of similarity between two textures can take a range of values.

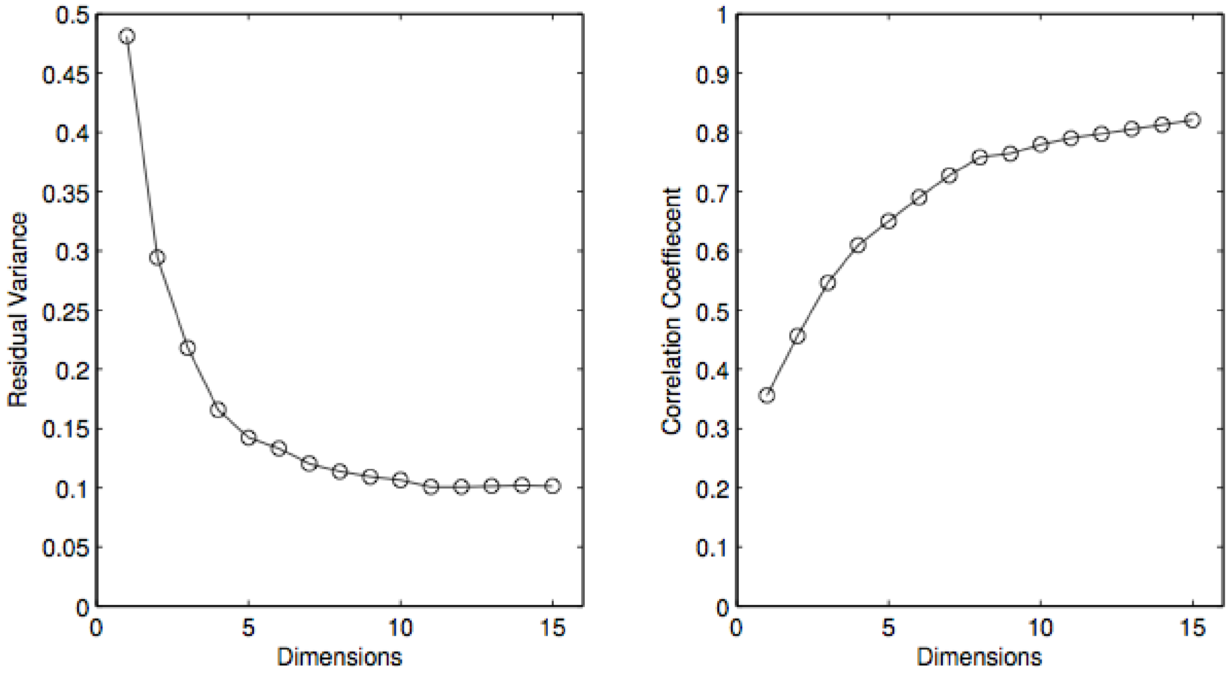


Figure 2: Isomap results. The correlations with the raw grouping data were computed using distances <= 1.

Predicting Perceptions: Proceedings of the 3rd International Conference on Appearance, pp. 17-20, Edinburgh, UK, ISBN 978-1-4716-6869-2, April 2012.

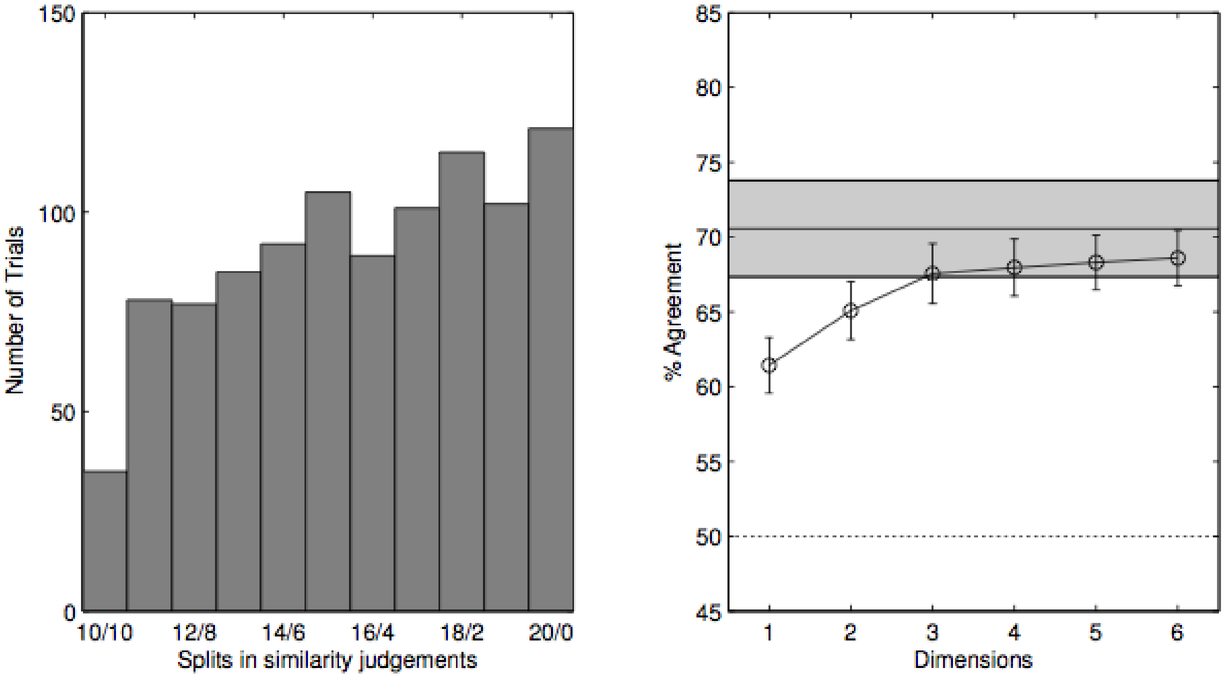


Figure 3: (left) Histogram showing the agreement between participants. At least fifteen of the twenty participants agreed with each other 63.3% of the trials. (right) Performance of the Isomap model as we increase the number of dimensions.