COMPARING COMPUTATIONAL AND HUMAN MEASURES OF VISUAL SIMILARITY

T. M. GALE*, Y. SUN, R. ADAMS, AND N. DAVEY

Neural Net Group,
University of Hertfordshire,
Hatfield, AL10, 9AB, United Kingdom
** QEII Hospital,*
Welwyn Garden City, United Kingdom
E-mail: t.gale, comrys, r.g.adams, n.davey@herts.ac.uk

There have been many attempts to quantify visual similarity within different categories of objects, which a view to using such measures to predict impaired recognition performance. Although many studies have linked measures of visual similarity to behavioral outcomes associated with object recognition, there has been little research on whether these measures are associated with human ratings of perceived similarity. In this work, we compare similarity measures extracted from Principal Component Analysis, Isometric Feature Mapping and wavelets representations with ratings of human subjects. Our results show that features extracted by calculating the standard deviation of wavelet coefficients provides the closest fit to the human rating data of all the methods we applied here.

1. Introduction

There have been many attempts to quantify the visual similarity of different objects and categories, and to use such measures to predict impaired recognition performance, most notably for category-specific agnosias. Different approaches have included contour overlap and shared features (Humphreys *et al.*, 1988), shape overlap measures derived from non-standardised images (Tranel *et al.*, 1997), Euclidean distance measures derived from standardised pixellated images (Laws and Gale, 2002a) and measures derived from unsupervised neural network representations (Gale *et al.*, 2001; Gale *et al.*, 2003). Humphreys and Riddoch (2002) suggest that Euclidean distance-based representations of pixellated images are not useful analogues for human perceptual representation because the latter does not proceed on the basis of such detailed information. Nonetheless, Euclidean distance-based measures of pixellated images do predict naming accuracy (Laws and Gale,

2002a) as well as naming latency to degraded visual input (Laws *et al.*, 2002). Laws and Gale (2002b) have argued that although such measures may not mirror the kind of processing that occurs in the human visual system, they nonetheless capture information about shape overlap, spatial orientation and distribution of shading and, as such, can provide a useful metrics of visual similarity.

Although many studies have attempted to associate measures of visual similarity with performance on object recognition tasks (e.g. picture naming), there is little research on whether these measures correlate with human judgements of visual similarity. In a recent study (Gale *et al.*, 2004) we compared similarity measures extracted from self-organising map (SOM) representations of greyscale pictures with ratings of perceived visual similarity provided by human participants for the same set of pictures. There was a strong statistical association between the 2 measures, suggesting that the clustering of representations within the SOM may be a useful model for human perceptual representation of certain object categories. In the current study, we compare our previous findings with some other computational measures of visual similarity to investigate whether these approaches can also provide useful techniques for modelling aspects of human perceptual categorisation.

2. Method 1: Data from Human Participants

2.1. *Stimuli*

Seventy basic-level categories (e.g. apple, dog, guitar, vase, airplane, etc.) were selected to represent a broad range of objects (35 living, 35 nonliving). Three-hundred-and-seventy different exemplar pictures were collected to represent these categories and the number of exemplars representing each basic level category varied between 4 and 7 (means for living: nonliving = 5.31: 5.28). Pictures were collected from online galleries and CD-ROM encyclopaedias and all chosen images depicted the referent item in a consistent and typical orientation (for example, see Figure 1). The pictures were standardized as follows: first, extraneous background material was carefully removed and the depicted items were normalized for orientation within each basic level category (for example, all examples of 'dog' were viewed side-on and facing the same way); the size of the pictures was then manipulated such that each depicted item fitted within a grid of 64 pixels square (4096 pixels in total) whereby the maximal dimension of the object touched the borders of the square; finally the images were converted to 8-

bit greyscale. Figure 1 displays exemplars for one basic level category and the full picture set is available on request.

Figure 1. Example standardized images for the basic-level category 'beetles'

2.2. *Human Ratings of Visual Similarity*

A group of 24 human subjects rated the degree of visual overlap (VO) for each of the 70 basic level categories. The same standardized pictures that were used to test the computational models were presented in their basic level categories and participants were asked to rate the level of VO in each of the 70 categories. All ratings were collected on a 5-point scale ranging from 1 'very similar' to 5 'very dissimilar'.

3. Method 2: Data from Computational Experiments

3.1. *Representations*

In this work, we consider 3 different computational representations, namely, Principal Component Analysis (PCA), Isometric Feature Mapping (ISOMAP) (Tenenbaum *et al.*, 2000) and wavelets (Vidakovic, 1999).

PCA is considered here because of its simplicity and wide domain of application. The motivation for the ISOMAP is to find meaningful low-dimensional structures hidden in high-dimensional observed data. Rather than calculating the Euclidean distances as in multidimensional scaling, this approach preserves the *geodesic distances* between all pairs of data points in the manifold (Tenenbaum *et al.*, 2000).

Recently, psychologists have applyed wavelets for modeling aspects of the visual system (Granlund, 1978; Daugman, 1980; Watson, 1983). Considering an image as a signal, it can be decomposed into wavelets of different sizes. Here a *2-D haar wavelet* transformation is employed to the pixel image, decomposing it into different directions: *vertical* (VD), *horizontal* (HD) and *diagonal* detail (DD) images of different resolutions. The representations are derived from second order statistics: the standard deviation of the coefficients at each level and each direction.

3.2. *Computational Experiments*

Each picture size is 64×64, 4096 pixels in total. In the current study, we calculate similarity using $s(\mathbf{x}, \mathbf{y}) = \frac{1}{1+d(\mathbf{x},\mathbf{y})}$, where $\mathbf{x}$ and $\mathbf{y}$ are representations for different images; d denotes Euclidean distance. In addition, we compare results with previous experimental results using a self-organising map (SOM) (Gale *et al.*, 2004).

4. Experimental Results

Results are displayed in Table 1. As shown, it suggests that the wavelet representations give better fit to the human rating data.

Table 1. Comparing computational and human measures of visual similarity. (P-value is the significance value.)

representations	corr. coeff	P-value
Pixel_based	0.44	0.0001
PCA	0.39	0.0008
ISOMAP	0.36	0.002
SOM	0.35	0.003
Wavelets	**0.56**	0.0001

5. Discussion

Our results show that features extracted by calculating the standard deviation of wavelet coefficients provides the closest fit to the human rating data of all the methods we applied here. The correlations between the wavelet representation and human ratings is considerably higher than that of the raw data and human ratings. Although these correlations are far from perfect, they are of an order of magnitude greater than those found for other predictor variables in visual object naming experiments (for example, see (Laws *et al.*, 2002)). In the next stage, we shall consider using a different distance measurement in computing similarity, rather than a simple Euclidean distance. We will undertake further work to investigate why wavelets are better predictors for the human data.

References

Daugman, J. G. 1980. Two-dimensional spectral analysis of cortical receptive field profiles. *Vision Research* **20**, 847–856.

Gale, T. M., N. Davey, K. R. Laws, M. Looms, and R. J. Frank 2004. Self-organising map representations of greyscale images reflect human similarity judgements. In *Proceedings IEEE IS*.

Gale, T. M., D. J. Done, and R. J. Frank 2001. Visual crowding and category-specific deficits for pictorial stimuli: a neural network model. *Cognitive Neuropsychology* **18**, 509–550.

Gale, T. M., K. R. Laws, R. Frank, and V. C. Leeson 2003. Basic level visual similarity and category specificity. *Brain and Cognition* **53**, 229–231.

Granlund, G. H. 1978. In Search of a General Picture Processing Operator. *Computer Graphics and Image Processing* **8** (2), 155–173.

Humphreys, G. W., J. Riddoch, and P. T. Quinlan 1988. Cascade processes in picture identification. *Cognitive Neuropsychology* (5), 67–103.

Humphreys, G. W. and M. J. Riddoch 2002. Do pixel-level analyses describe psychological perceptual similarity? A comment on 'Category Specific Naming and the visual characteristics of line-drawn stimuli' by Laws and Gale. *Cortex* **38**, 3–5.

Laws, K. R. and T. M. Gale 2002a. Category-specific naming and the 'visual' characteristics of line drawn stimuli. *Cortex* **38**, 7–21.

Laws, K. R. and T. M. Gale 2002b. Why are our similarities so different? A reply to Humphreys and Riddoch. *Cortex* **38**, 643–650.

Laws, K. R., V. C. Leeson, and T. M. Gale 2002. The effect of 'masking' on picture naming latencies. *Cortex* **38**, 137–147.

Tenenbaum, J. B., V. d. Silva, and J. C. Langford 2000. A global geometric framework for nonlinear dimensionality reduction. *Science* **290**, 2319–2323.

Tranel, D., C. G. Logan, R. J. Frank, and A. R. Damasio 1997, Oct. Explaining category-related effects in the retrieval of conceptual and lexical knowledge for concrete entities: operationalization and analysis of factors. *Neuropsychologia.* **35** (10), 1329–1339.

Vidakovic, B. 1999. *Statistical Modeling by Wavelets*. John Wiley & Sons, Inc.

Watson, A. B. 1983. Detection and recognition of simple spatial forms. In *Physical & Biological Processing of Images*, pp. 100–114. Berlin: Springer-verlog.