Self-Organising Map Representations Of Greyscale Images Reflect Human Similarity Judgements

Tim M. Gale, Neil Davey, Keith R. Laws, Martin Loomes, Ray J. Frank

Abstract— **In this study we assessed a Kohonen network's ability to represent visual similarity between grayscale pictures and whether these representations were associated with human ratings of perceived similarity. We trained a Kohonen network (SOM) with 370 standardized grayscale pictures deriving from 70 basic level object categories (e.g. dog, apple, chair, etc.) and measured, for each category, the average euclidean distance of the SOM output patterns to provide an index of the visual similarity between exemplars of the same basic level category. We then asked human subjects to provide visual similarity ratings for the same categories of stimuli and compared these with the measures extracted from the SOM. The significant correlation between the SOM and human measures suggests that a SOM may be a useful way of modeling certain stages of human visual categorization. Interestingly, the human ratings showed category-specific differences in the level of similarity ascribed to living and non-living things. However, this pattern was not reflected in the SOM representations of the same stimuli. This has important implications for theories of object recognition and, specifically, our understanding of category-specific naming impairments.**

Index Terms—**Self Organizing Map, Visual Object Recognition, Similarity, Category-specific disorder**

I. Introduction

Many studies of visual object recognition have used, or made reference to, neural network models (e.g. [1-3]). In many cases these models articulate pre-existing theories and, although they may sometimes predict human performance, there is little evidence of any association between the kind of representations that develop in the models and those that may be important in perceptual processing. For example, the majority of neural network models of visual object recognition are trained with vectors of artificially coded 'visual' attributes (e.g. 'is brown', 'is tall', 'has legs', etc.) that are somewhat unlikely to reflect properties of real sensory stimuli. Moreover, the vast majority of models to date have employed supervised learning algorithms and these do not reflect the kind of learning that occurs in early perceptual development. This is an important issue because the acquisition of certain object categories and concepts precedes the development of linguistic ability during infancy so it is unlikely that 'target states' are present during early categorical learning (see [4]). For this reason, we have previously argued that self-organizing neural networks are better suited to modeling pre-semantic stages of visual processing than supervised networks [3].

In a previous study [3] we presented a Kohonen network (SOM) with a range of standardized grayscale object images in an attempt to model the categorization of visual stimuli in a less artificial way. These stimuli derived from a range of categories and sub-categories, each with multiple exemplars. Rather than concentrating specifically on the winning unit activation, we measured the pattern of activation across the whole of the output map (following [5]). One aspect of this work was to explore whether the strength of clustering in the SOM representations varied across different object categories (e.g. animals vs. clothing vs. musical instruments, etc.). This holds particular theoretical interest within the study of category-specific recognition disorders. Such disorders usually present following certain types of brain injury with patients typically showing degraded object naming performance for certain classes of item (e.g. living things), in contrast to preserved object naming for other classes (e.g. nonliving things). Such cases have been highly influential in the development of theories and models relating to visual object processing and knowledge representation in the human brain. More recently, however, such theories have placed greater emphasis on properties of real-world objects and stimuli which might influence our ability to differentiate and name them, especially under processing constraints (for example, brain injury or degraded viewing conditions). One such property, that has been the subject of much recent debate, is visual crowding (see [2, 3, 6-8]). The visual crowding hypothesis suggests that exemplars from certain object categories are more visually similar to each other. That is to say that there is greater 'perceptual' overlap between examplars from certain categories (typically living things; [8]) than others. In visually

Tim Gale is a visiting fellow at The University of Hertfordshire and is with the Department of Psychiatry, QEII Hospital, Howlands, Welwyn Garden City, Herts, AL7 4HL, UK (telephone: +44 (0)1707 369058, e-mail: t.gale@herts.ac.uk)

Neil Davey is a lecturer with the Department of Computer Science, University of Hertfordshire, Hatfield, Herts, UK (e-mail: r.n.davey@herts.ac.uk)

Keith Laws is a reader with the Brain and Cognition Group, Nottinghman Trent University, Nottingham, UK (e-mail: keith.laws@ntu.ac.uk)

Martin Loomes is a Professor in the department of Computer Science, University of Hertfordshire, Hatfield Herts, UK (email: M.J.Loomes@herts.ac.uk)

Ray Frank is a lecturer with the department of Computer Science, University of Hertfordshire, Hatfield, Herts, UK (e-mail: r.j.frank@herts.ac.uk)

0-7803-8278-1/04/$20.00 ©2004 IEEE

crowded categories, exemplar discrimination would theoretically be more difficult and this may have a profound impact on the ease with which category exemplars are named. In neurologically intact individuals, such an impact may only be detectable by very sensitive tests such as response latency to object naming, while for brain-injured individuals, it may be strongly evident in naming ability per se.

To date, most work on visual crowding has focused on similarity at the 'superordinate' categorical level (e.g. how similar are examples of animals? or how similar are examples of vegetables?). However, in day-to-day object recognition, it is the *basic level* of categorization that is considered fundamentally important. The basic level of categorization (e.g. dog, cat, chair, guitar, shirt etc.) is much less inclusive than the superordinate level. It is considered the most salient level of categorization for humans because it is purported to reflect the greatest level of *within-category similarity* and *between-category dissimilarity*, thereby making object identification easiest at this level (for example, it is easier to categorize as 'bird' than 'lapwing' and it is more useful to categorize as 'bird' than 'animal'). Previous studies have also underscored the importance of the basic level, particularly in early conceptual development and language (e.g. [9, 10]). However, to date, there has been no attempt to quantify similarity within basic level categories and to see whether this has any association with human behavioral data.

In the current study we investigate the representation and coherence of basic level object categories within a Kohonen network model and, importantly, whether these are representations are associated with human ratings of perceived similarity. We use a novel set of standardized grayscale object images and compare the measures of similarity deriving from the neural network representations with ratings of visual similarity provided by human subjects for the same set of pictures. This work addresses two important issues, namely, (i) do the neural network representations of our pictures predict behavioral responses to the same pictures and (ii) is there any basis for assuming greater visual crowding for living things at the basic level, which might influence object naming and/or basic level categorization under normality and pathology?

II. METHOD

A. 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 and 5.28 respectively). Pictures were collected from online image galleries and CD-ROM encyclopaedias and all chosen images depicted the referent item in a consistent and typical orientation (for example, see fig. 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 pixel grid (following [3, 6]); 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 from the first author.

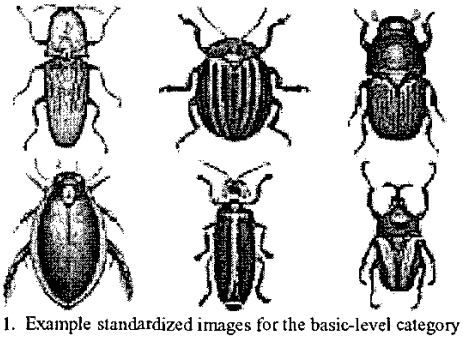

Fig. 1. Example standardized images for the basic-level category 'beetle'

B. The Model

We implemented a standard Kohonen map ([11-13]) also known as a self-organizing map (SOM). This is an unsupervised network that uses a competitive learning algorithm to develop reduced-dimensional, topology-preserving, representations of input patterns. An almost identical model is described in [3] so just a brief overview is presented here. Our model comprised a 4096-dimensional (i.e. 64 by 64 pixel) input vector (A) that was fully connected to a SOM of 7 units square (see figure 2). So, each output unit (oi) in the SOM was connected to the input vector (A) with a 4096-dimensional weight vector (Wi). At each iteration, the activation values across the SOM were derived by comparing the Euclidean Distance between the input vector (A) and the weight vector (Wi) for each of the 49 output units. The output map was calculated using a function that returned a value in the range of 0-64 where 64 represented a unit whose weight vector was as far away as possible from the input vector and 0 represented a unit whose weight vector was identical. The output unit with the lowest Euclidean distance from the input vector was regarded as the 'winner'. The training rule updated the weights of the winning unit, and those of a neighborhood surrounding the winner, moving them closer in value to the input vector. The initial neighborhood size was 7 units square (i.e. the 'hamming' distance was 3 units either side of each winning unit and the SOM surface was wrapped around) although this decreased to zero, linearly, over training time. The update rule for the neighborhood was defined by a non-linear function (defined by 1/hamming distance) such that the weight vectors of units close to the winner were modulated to a greater extent by the input vector than those that were further away. It is these

correlated zones of activation that permit self-organization (see also [5]).

The focus of interest with regard to pattern classification in SOMs is usually on 'winning' units because these will typically identify different prototypes or groups of patterns. However, in these experiments, we were also interested in the distribution of activation across the other output units and this representation can be conceived as a contour map. The 'winning unit' (i.e. the most highly activated) is important in this model but does not have exclusive diagnosticity (i.e. it is possible for 2 training patterns to activate the same winning unit, yet have different activation distributions across the remaining units). SOMs have been used to model tasks that require perceptual (i.e. bottom-up) categorization (e.g. [14, 15]), because they develop representations in the absence of top-down constraints. Thus, the final output states of the network are self-evolving rather than pre-determined.

Ten SOMs (each randomly configured) were trained with the set of 370 images. Thes stimuli were presented in random order for 2000 epochs and the learning rate was set initially at 0.05 but decreased linearly over training time. At completion, the output representation for every image was recorded as a 49-dimensional vector (see fig. 2). These vectors were compared with each other to extract a matrix of euclidean distance scores between every possible pair of SOM representations (370 by 370). Scores were then averaged across related exemplars (i.e. within each basic level category) to provide an average measure of overlap for each of the 70 basic level categories. These averages were standardized as proportions of the maximum possible Euclidean Distance score (448 in this case) and, finally, this was averaged across all 10 replications.

A low mean euclidean distance score between category exemplars would indicate a high level of overlap within a given category, whereas a high euclidean distance score would indicate that the category members were visually dissimilar to each other.

C. Human Ratings of Visual Similarity

A group of human subjects rated the degree of visual overlap (VO) for each of the 70 basic level categories. Twenty-four (12 male, 12 female) provided ratings of the extent to which all items with the name, e.g. 'beetle' looked similar (these participants did not see the actual sets of stimuli used to train the SOM and their ratings were termed 'VO_verbal'). Another 24 subjects (10 male, 14 female) saw exactly the same standardized pictures that were presented to the SOM. These were presented in their basic level groupings and participants were asked to rate the level of VO within each of the 70 basic level categories. These ratings were termed 'VO_visual'. All ratings were collected on a 5-point scale ranging from 1 'very similar' to 5 'very dissimilar'.

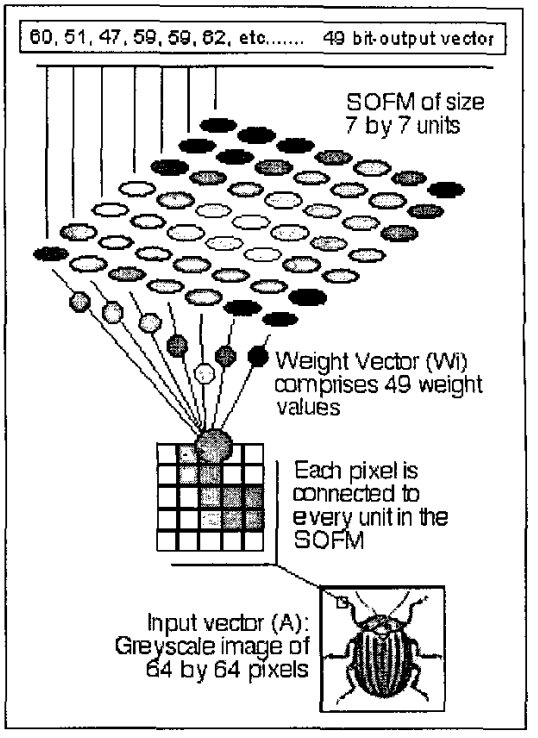

Fig. 2. An overview of image processing within the model

III. Results

A. Statistical Association Between Similarity Measures

VO_visual and VO_verbal were strongly correlated (r = 0.53, $p < 0.0001$) and this association remained significant when living and nonliving things were analyzed separately (living: r = 0.37; p = 0.03; nonliving: r = 0.53, p = 0.001). Under VO_verbal, participants were rating the similarity of imagined exemplars, so it is interesting that this has a strong association with ratings made to specific sets of exemplars chosen by the experimenters. The measures derived from the SOM correlated significantly with the human ratings: (for VO_visual, r = -0.35, p = 0.003). Looking again at living and nonliving things separately, both had significant correlations (living: r = -0.41, p = 0.02; nonliving r = -0.36, p = 0.03). The mean SOM output euclidean distance values also correlated with VO_verbal but a different pattern emerged for living and nonliving things: r = 0.16, p = 0.36 NS; nonliving: r = 0.59, p = 0.002). 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 [16]).

B. Category-specific Differences

There was no significant difference between living and nonliving things in terms of mean euclidean distance for SOM representations of basic level categories (means = 0.125 vs. 0.127 respectively, $F < 1$, NS). However, For VR_verbal, living things were rated as having greater within-category (basic level) similarity than nonliving things (3.7 vs. 3.3: $t_{1,\ 68} = 3.2$, p = 0.002). For VR_visual,

living things were also perceived to have higher within-category similarity (3.6 vs. 3: $t_{1,\ 68} = 4$, $p < 0.0001$). Thus, although the SOM measures did not reveal any category-specific differences, such differences were evident in the ratings, despite the fact that these measures actually correlated highly with the SOM measures.

IV. Discussion

The significant correlation between the SOM measures and VO suggests that a SOM may be a useful tool for modeling the acquisition of visual categories. In our model, the SOM reduced the dimensionality of each input pattern from 4096 down to 49 and our results suggest that the data which is preserved in these representations, may be reflective of the kind of information that could be used in human perceptual categorization. Further work will involve manipulating the size of the SOM to see whether this has an effect on the strength of correlation observed.

Living and nonliving things differed in the strength of association between average SOM euclidean distance (for basic level categories) and VO_verbal (but not VO_visual): while living things showed a weak association ($r = 0.16$), there was a much stronger relationship for nonliving things ($r = 0.59$). Looking at this further, a notable split emerged within subcategories of living things; while the SOM representations of animals, insects and sea creatures all indicated reasonably high overlap, those of fruit and vegetables did not. Indeed, if fruit and vegetables are excluded from analyses, the correlation between the average SOM euclidean distances and VO_verbal for living things increases to a magnitude more similar to that observed for nonliving things ($r = 0.44$, $p = 0.03$). One possible explanation for the discrepancy between fruits/vegetables and other living thing categories is the modulatory effect of color. Although different examples of the same type of fruit/vegetables (e.g. compare different examples of: strawberry; parsnip; carrot; lettuce; tomato) can have quite different shapes, their color tends to be less variant and more diagnostic (see also [17]). Color may therefore affect VO_verbal ratings disproportionately for these categories, increasing the perceived amount of basic-level visual similarity. Indeed, color is rated as being much more critical for recognizing fruits and vegetables than other (both living and nonliving) categories (see [18]), and this would not be captured by these simulations because the pictures were all gray. In short, the SOM may capture perceived visual similarity quite well when the principal determinant is pixel luminance (and hence, indirectly, shape/detail). However, if factors other than these exert an influence on human judgements, this simple model has no mechanism for reflecting this. This is supported by the fact that a living/nonliving difference was not observed in the correlations between the SOM measures and VO_visual scores, where the subject ratings were made to standardized grayscale stimuli rather than imagined examples. So, as a model of perceptual categorization, the SOM may fail to capture reported basic level similarity relationships for certain categories simply because it cannot process color information. Nevertheless, for all other categories, there is a statistically significant association between the SOM measures and VO (both verbal and visual), suggesting that the model represents categorical boundaries in a way that accords with subjective reports of basic level visual similarity.

There was no difference in the extent of basic level overlap recorded by the SOM for living and nonliving categories. It would therefore be reasonable to conclude that there are no physical characteristics within the pictures which predispose living things to be more visually crowded than nonliving things. However, human subjects rated living things as having greater visual overlap, irrespective of whether their judgements were based on real or imagined exemplars. This tendency has also been reported previously (see [19, 20]), though this is the first study to compare human ratings with more objective measures (i.e. SOM representations). So, on what basis does this pattern arise?

This dissociation raises an important issue that has been overlooked in previous research. It hints at the importance of distinguishing between similarity that exists for real-world objects or pictures (what we might call 'assembled' similarity) and similarity that exists within stored human representations (what we might call 'addressed' similarity). Previous studies of visual crowding have not made such a distinction and there has been an unspoken assumption that stored structural representations simply capture the properties of assembled measures. That a living/nonliving difference was observed in our measure of 'addressed' similarity, but not in the 'assembled' measures, suggests that visual categorization may exert a disproportionately stronger clustering effect on living things (i.e. living things have more coherent visual prototypes) [1]. Nevertheless, this must arise via an interaction of top-down and bottom-up processing since no living/nonliving difference is evident at 'input' level.

In conclusion, this work suggests that if visual crowding effects play a role in visual categorization and object recognition, these effects may well be independent of the stimuli themselves. Thus, it is possible that stored structural object representations in the brain have evolved in a way that clusters living things more tightly than nonliving things. However, this is likely to arise through an interaction of cognitive and perceptual processes and cannot be visual crowding per se.

Acknowledgements

The authors are grateful to Thanusha Sivakumaran and Jenny Hayes for assistance in collecting similarity ratings.

References

[1] M. FARAH and J. L. MCCLELLAND, "A computational model of semantic memory impairment: modality specificity and emergent category specificity," *Journal of Experimental Psychology: General*, pp. 339-357, 1991.

[2] G. W. HUMPHREYS, C. LAMOTE, and T. LLOYD-JONES, "An interactive activation approach to object processing: effects of structural similarity, name frequency and task in normality and pathology.," *Memory*, pp. 509-550, 1995.

[1] Although it is possible that subject ratings are influenced by the *belief* that living things are more similar and nonliving things less similar (rather than actually retrieving and evaluating stored representations or exemplars).

[3] T. M. GALE, D. J. DONE, and R. J. FRANK, "Visual crowding and category-specific deficits for pictorial stimuli: a neural network model," *Cognitive Neuropsychology*, pp. 509-550, 2001.
[4] P. C. QUINN and P. D. EIMAS, "On categorization in early infancy. Merill-Palmer Quarterly, 32, 331-363, 1986.," *Merill-Palmer Quarterly*, pp. 331-363, 1986.
[5] P. G. SCHYNNS, " A modular neural network model of concept acquisition," *Cognitive Science*, pp. 461–508, 1991.
[6] K. R. LAWS and T. M. GALE, "Category-specific naming and the 'visual' characteristics of line drawn stimuli.," *Cortex*, pp. 7-21, 2002.
[7] K. R. LAWS and T. M. GALE, "Why are our similarities so different? A reply to Humphreys and Riddoch," *Cortex*, pp. 643-650, 2002.
[8] G. W. HUMPHREYS, J. RIDDOCH, and P. T. QUINLAN, "Cascade processes in picture identification," *Cognitive Neuropsychology*, pp. 67-103, 1988.
[9] E. ROSCH, C. B. MERVIS, W. D. GRAY, D. M. JOHNSON, and P. BOYES-BRAEM, " Basic objects in natural categories," *Cognitive Psychology*, pp. 382-439, 1976.
[10] C. B. MERVIS and M. A. CRISAFI, "Order of acquisition of subordinate-, basic-, and superordinate-level categories. Child Development," *Child Development*, pp. 258-266, 1982.
[11] T. KOHONEN, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, pp. 56-69, 1982.
[12] T. KOHONEN, "Clustering, taxonomy and topological maps of patterns," presented at Sixth International Conference on Pattern Recognition, Silver Spring, MD, 1982.
[13] T. KOHONEN, *Self-Organization and Associative Memory*: Berlin: Springer-Verlag, 1988.
[14] S. LAWRENCE, C. L. GILES, A. C. TSOI, and A. BACK, "Face recognition: a convolutional neural network approach," *IEEE Transactions on Neural Networks*, vol. 8, pp. 98-113, 1997.
[15] A. J. LUCKMAN, N. M. ALLISON, A. W. ELLIS, and B. M. FLUDE, "Familiar face recognition: a comparative study of a connectionist model and human performance," *Neurocomputing*, pp. 3-27, 1995.
[16] K. R. LAWS, V. C. LEESON, and T. M. GALE, " The effect of masking on picture naming," *Cortex*, pp. 137-147, 2002.
[17] C. J. PRICE and G. W. HUMPHREYS, "The effects of surface detail on object categorization and naming," *Quarterly Journal of Experimental Psychology*, pp. 797-828, 1989.
[18] K. R. LAWS and S. AKHTAR, "Naming feature centrality: an advantage for living things," *Brain and Cognition*, vol. 53, 2003.
[19] K. R. LAWS and C. Neve, " A 'normal' category-specific advantage for naming living things.," *Neuropsychologia*, pp. 137-147, 1999.
[20] O. H. TURNBULL and K. R. LAWS, "Loss of stored knowledge of object structure: implications for 'category-specific' deficits," *Cognitive Neuropsychology*, pp. 365–389, 2000.