

COMPUTATIONAL AND COGNITIVE STUDIES IN SIMILARITY

Evaluating a Neural Network Model of Visuospatial Similarity in Design

JULIE R JUPP AND JOHN S GERO
University of Sydney, Australia

Abstract. In this paper the cognitive plausibility of utilising self-organising maps in similarity assessments of 2D design diagrams is analysed using human-subject experiments which focus on designers. The experiments reported here are designed to address the validity of the computational approach to similarity and investigate feature types, feature salience and the role of context in cognitive assessments.

1. Introduction

This paper presents a cognitive evaluation of a formal model of 2D visuospatial similarity assessment in design and discusses validation aimed at testing the model's utility. In a previous paper (Jupp and Gero 2005) this research presented a computational model of similarity that compares qualitative re-representations of 2D design diagrams. A model of 2D design similarity is only useful if it can provide results that can be recognised or matched by an observer's judgements. Cognitive validation is explored here in studies of intuitive similarity assessments made by designers and compares these judgments against the results of model.

It is expected that the results from cognitive similarity assessments will mirror most of the distinctions made by the computational model, or at least identify correlations and insights into the clusters and ordering of 2D design diagrams made by the model and by designers. It is also expected that studying human-subject similarity assessments will provide some insight into the different types of perceptual grouping strategies as well as the different types of features used during comparison making and their relative salience using different similarity-based assessment methods.

The remainder of this paper is divided into six sections. The approach to similarity assessment is presented in Section 2 in conjunction with a summary of the computational model presented by Jupp and Gero (2005) and the general results obtained. The section also contains a discussion of

research in psychological similarity assessment outside the design domain. The experiment design and set-up are described in Section 3. Three experiments are presented using 36 architectural design students from the 1st and 5th years. The results are presented in Section 4 with an analysis that tests five hypotheses. Section 5 discusses these results in relation to other studies of similarity assessment contained in the literature. Section 6 then concludes the paper.

2. Similarity Assessment

The approach to similarity assessment adopted by this research moves away from the idea of similarity as the outcome of a direct comparison procedure and towards the idea of similarity as a process whose outcome can only be reported in a post-hoc fashion. The neural network approach suggests that similarity is related to the way in which information is processed, where the reporting of similarity is a meta-cognitive process requiring the explicit comparison of information prior and subsequent to processing by the cognitive system (Thomas and Mareschal 2000).

2.1 COMPUTATIONAL SIMILARITY ASSESSMENT

The neural network model of similarity previously proposed (Jupp and Gero 2005), called Q-SOM, is based on a Self-Organising Map (SOM) and a hierarchical feature-based schema for qualitative re-representation of visuospatial attributes. The features derived from the re-representation of 2D design diagrams are capable of describing three levels of spatial information, namely: morphology, topology and mereotopology. The similarity assessments made by the Q-SOM model relies on these features as input to the network. Processing follows four consecutive stages: (i) recognition, extraction and encoding of different levels of spatial attributes; (ii) initial feature selection of spatial attributes; (iii) categorisation via unsupervised learning of 2D diagrams based on available features; and (iv) identification of clusters using K-Means clustering.

Q-SOM was tested on a corpus of 61 plan diagrams from the 20th century architect Frank Lloyd Wright (Jupp 2005). Wright's residential plan designs from 1885 to 1940 were used as the dataset for these experiments. The designs from Wright's corpus are structured (agreed upon by both historians and critiques) into four periods: early, prairie, transition and usonian. From each of the four periods, a fixed ratio of approximately 3.5:1 determined the number of diagrams randomly selected for training, giving a total of 15 diagrams. The training set comprised: 2 early, 6 prairie, 2 transition and 5 usonian diagrams. Networks were trained unsupervised and no explicit contextual information was input. A level of contextual information could be said to be implicit in the feature semantics derived from diagrams. There is

no explicit connection with a design context, for example any kind of design scenario or design requirement.

The output of Q-SOM's networks were evaluated statistically using techniques from conventional text-based analysis including: Precision (Slonim et al 2002), the Jaccard method (Downton and Brennon 1980), and the Fowlkes-Mallows method (Fowlkes and Mallows 1983). Based on these internal performance measures it was observed that using multiple classes of features yields better results for classifying diagrams according to the four periods of Wright's work than any one single dimension. Results revealed that using a larger set of features based on a combination of available morphological, topological and mereotopological features yields better results than using a smaller subset of features. For these results to be meaningful to designers, they must now be evaluated in relation to cognitive similarity assessments. Validating the Q-SOM's correspondence to human-subject similarity judgments can be achieved by demonstrating that the outputs of the SOMs are similar to those judged by an observer.

2.2 PSYCHOLOGICAL SIMILARITY ASSESSMENT

Assessing the similarity of objects is neither simple nor well understood (Palmeri and Gauthier 2004). How mental representations of categories work and how they are related to similarity judgements remains a controversial question (Lambert and Shanks 1997, Medin et al 2000, Palmeri and Gauthier 2004). There are four major psychological models of similarity: geometric, featural, alignment-based, and transformational; and all four approaches have enjoyed some success in quantitatively predicting people's similarity assessments. The approach of the Q-SOM model summarised in the previous section, and which is now tested here, is featural.

There is great variability within the literature regarding the types and combinations of techniques used to validate neural networks. Although computational evaluation is widely used to validate the output of neural networks, (e.g. the Precision, and the Jaccard and Fowlkes-Mallows methods) cognitive validation is less prevalent. This is most likely due to the complexities and ambiguities in studying human-subjects since, in general, similarity based assessment may not be a unitary measure and may also depend on representations that are constructed and changed by the designer during comparison-making. Goldstone (1994) enumerates the difficulties in studying similarity assessments, reporting that the explanatory role of similarity may not be a unitary phenomenon. Similarity can be influenced by context, perspective, choice, alternatives, and expertise (Medin et al. 1993, Tversky 1977). Different processes for assessing similarity are probably used for different design tasks, and design diagrams. For example, the similarity of diagrams can be assessed based on its complexity or can be

influenced by the design task at hand. Specific features may be selectively weighted during assessments and evidence from previous research indicates that the weighting of features in similarity judgments may also vary dynamically during processing (Goldstone 1994).

In light of the difficulties in studying cognitive similarity assessments, one of our principal concerns in designing experiments lies in studying a variety of methods for similarity assessment and ensuring that the assessments made are intuitive and not guided or constrained by strict definitions. All participants shared a common understanding of the concept of similarity and the type of perceptual information being assessed. The following section describes the experiment design.

3. Experiment Design

The main objective of the cognitive experiments is to evaluate the correspondence between the subjects themselves and between subjects and the Q-SOM's output rather than the four periods defined by historians and critiques which partition Wright's designs into some kind of category or style. Wright's periods are defined as the 'correct' classification and the evaluation of results are based on these characterisations.

The experiment consisted of three questions. The same 15 diagrams used to train the Q-SOMs networks (Section 2.1) were again utilised as this study's design corpus. As in the computational experiments, the selection of Wright's residential work includes diagrams from the early, prairie, transition and usonian periods.

3.1 SUBJECTS

Participants were recruited from the 1st and 5th (final) years of the Bachelor of Design Architecture degree at Sydney University. The 36 participants were all unpaid volunteers. Subjects' ages ranged from 18 to 32 years. The 36 subjects included 21 females and 15 males. For approximately 75% of students English is their first language and the remaining 25% of students were fluent in English. Subjects responded to each question at the same time, i.e., in group sessions, and were allocated the same amount of time to provide their answers.

3.2 SET-UP

The selection of 15 plan diagrams is illustrated in Figure 1 and Table 1 lists corresponding label and period, where E = early, P = prairie, T = transition and U = usonian.

The general method of the experiments enabled participants to study and then rank the corpus. Before subjects undertook each experiment, a training session was provided to familiarise designers with the task. During this

training period general definitions and guidelines were described, including: the concepts of physicality, and similarity. In addition, illustrations of diagrams were controlled using strict criteria.

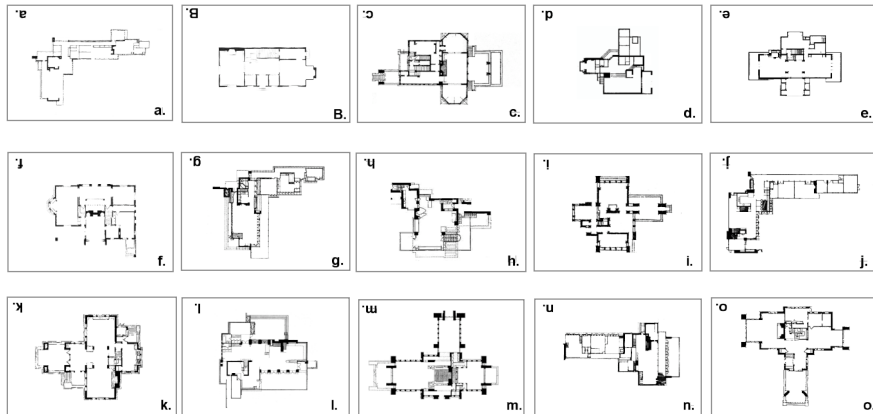


Figure 1. Selection of 15 plan diagrams for the design corpus of Frank Lloyd Wright

TABLE 1. Correct partitioning and ordering of diagrams based on Wright’s periods.

LABEL	b	f	c	k	i	o	e	m	d	h	g	j	a	n	l
DIAGRAM	CHARNLEY HOUSE 1891	WINSLOW HOUSE 1898	HENDERSON HOUSE 1901	BARTON HOUSE 1903	COMM. HOUSE 1903	LITTLE HOUSE 1903	DERHODES 1906	ULLMAN (UNBUILT)	YOUNG HOUSE 1929	KAUFMANN HOUSE 1935	JACOBS HOUSE 1936	LUCK HOUSE 1936	GARRISON HOUSE 1939	NEWMAN 1939	SCHWARTZ HOUSE 1939
PERIOD	E	E	P	P	P	P	P	P	T	T	U	U	U	U	U
TOTAL N ^o .	2		6						2		5				

3.2.1 Understanding of Physical Comparison

Subjects were instructed to base assessment on a diagram’s physicality and asked to identify similarities based only on visuospatial information. Subjects were provided with a general definition of physicality and visuospatial information:

- Material elements or characteristics such as shapes and shape elements
- Spatial arrangements and relationships perceptible through vision

Subjects were also provided with criteria which should not be considered as physical or visuospatial information, including:

- Orientation – subjects were able to rotate diagrams
- Overall Scale – diagrams were scaled to the approximate scale
- Contrast, line weight and quality – diagrams were uniformly printed

3.2.2 Understanding of Similarity

To ensure subjects shared the same general understanding of similarity, the following simple definition was provided:

- Corresponding physical elements producing some visual resemblance
- Shared physical characteristics of two or more diagrams

3.2.3 Diagram Illustrations

The experimental procedure represented Wright's plan diagrams uniformly on A5 cards. Each diagram was labelled from "a" to "o" and to ensure diagrams were presented in a consistent manner and to reduce any bias created by orientation, all A5 cards were labelled on the reverse-reflected corners. Plan diagrams were scaled to approximately the same scale, and were printed with the same line quality and contrast.

The objective of these criteria and definitions are to ensure that the responses obtained from subjects provide data from which to accurately compare and analyse any interrelationships which might exist.

3.3 RANKING TASKS

To study similarity assessments of design diagrams, three types of experiments were undertaken, which all required subjects to rank the 15 diagrams, namely: a simple complexity-based ranking task, a target-based ranking task and a contextually dependent ranking task. Following each training period, each task was read aloud to subjects by the instructor (the author) as well as being displayed to subjects on a projector and their own individual computer screen as shown in Figure 2(a). The tasks performed by subjects in each experiment consisted of two separate phases.

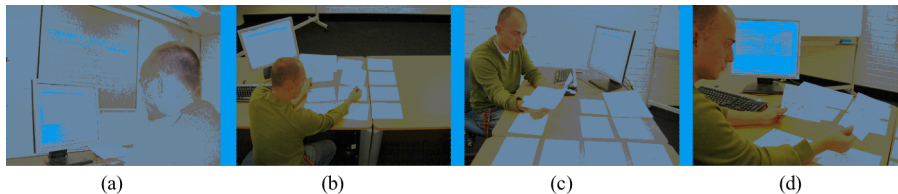


Figure 2. Experiment set up and testing environment showing: (a) Projected instructions, (b) and (c) sorting area and layout space, and (d) individual results input monitor.

In phase 1 participants were given eight minutes in which to study and assess the similarities of the corpus. The 15 randomly ordered diagrams were given to subjects in a closed envelope. Participants were able to arrange the corpus on the layout space provided, Figure 2(b) and (c). Once the set time had lapsed, subjects were instructed to stop their assessment and commence phase 2 of the experiment.

In phase 2, designers were given eight minutes, in which to complete a spreadsheet and specify their responses. Once subjects' completed phase 2 they were then asked to describe those attributes used to rank diagrams. Based on the responses obtained from all three experiments, subjects' descriptions were divided into five general categories:

- individual shape type,
- feature types contained within individual shape type,
- arrangement or relationships of individual shape type,
- overall or bounding shape type,
- combination of the above, and
- unknown.

The following sections present the experiments and questions in further detail.

3.3.1 Experiment 1

The first question asks subjects to rank design diagrams from highest to lowest according to their complexity, where diagrams can be ranked equally.

Question 1 does not rely on any explicit contextual information and is defined as a 'default' ranking case since there is no design context and the assessment criteria is simply the physicality of the 2D diagrams themselves.

3.3.2 Experiment 2

Question 2 asks subjects to rank design diagrams from highest to lowest according to their similarity to a target diagram where diagrams can be ranked equally. The target diagram is also from Wright's design corpus: the Pope House; and belongs to the usonian period as shown in Figure 3.

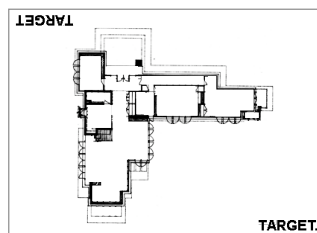


Figure 3. Target: the Pope House, 1940 Frank Lloyd Wright

3.3.3 Experiment 3

The final question specifies an explicit design context by defining a design task. This question utilises the 15 diagrams as a precedent library, which is to be considered by subjects prior to undertaking a design task. Using the brief provided, subjects must interpret the design requirements and assess the similarity of the 'precedent' diagrams in relation to the existing design, shown in Figure 4, and their interpretation of the design brief (see boxed text).

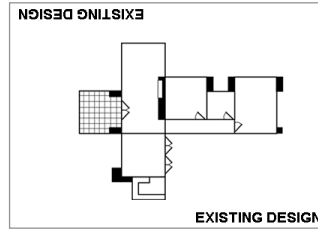


Figure 4. Existing sketch design used in design task.

Using the existing design diagram and the design brief subjects were asked to rank diagrams from highest to lowest according to their similarity to the design task.

Design Brief: Alterations and additions of an existing residential design are required to increase sleeping and living spaces, according to the following specifications of building layout:

- additional sleeping areas to accommodate two children;
- larger lounge, dining and kitchen areas; and
- outdoor living area;

Before undertaking this task a corpus of architectural plan designs is provided as precedents to assist you in designing.

4. Subjects' Responses and Analysis

No significant differences were found in subject responses based on the participant's level of design education, i.e., 1st (novice) or final year (expert) students. There was also no significant evidence found for differences based on gender or language. The following analysis of results considers the total of responses for each experiment. Tied ranks were normalized by the mean of the ranks for which they tie, assuming the number of ranks is equal to the number of diagrams compared.

4.1 TESTING HYPOTHESES

The first three hypotheses tested compare the responses among the subjects themselves and the last two hypotheses compare subjects' answers in relation to the results of the Q-SOM model.

4.1.1 Hypothesis 1: Subjects' similarity assessments will show correlation.

The first hypothesis is tested with Kendall's coefficient of concordance W for multiple rankings (Daniel 1978). Here, this test uses the normalised

subjects' responses such that each design diagram in each question has 15 different ranks.

The test statistic W for each question in Experiments 1 to 3 is shown in Table 2. For Question 1 (complexity-based) and Question 2 (target-based) the value of W is relatively high with 0.53 and 0.75 respectively. Based on these results, the hypothesis can be accepted for responses to Questions 1 and 2 with a Type I probability equal to 0.13. The relatively large number of subject responses makes the test statistically significant.

The value of W for Question 3 (contextually dependent) is very low (0.21), which means there is little agreement amongst subjects. It is inconclusive as to whether this hypothesis can be accepted under the constraints of Question 3.

TABLE 2 Kendall's coefficient for Experiment Questions 1, 2 and 3.

QUESTION	EXPERIMENT 1 TO 3		
	1	2	3
W	0.53	0.75	0.21
p	< 0.66	< 0.85	< 0.98

The standard deviation of normalised ranks for each question can also provide an indication of whether the subjects' responses are more associated with particular ranks. It was expected that observers would agree on diagrams deemed to have higher similarity as well as those diagrams with lower similarity, i.e., first and last rankings, but there would be higher levels of discrepancy in the responses of those diagrams ranked in-between. Figure 5 shows that the agreement across Questions 1 and 2 does clearly follow this pattern, i.e., for the first four ranks and the last two to three ranks and is illustrated by the arc indicating higher levels of agreement.

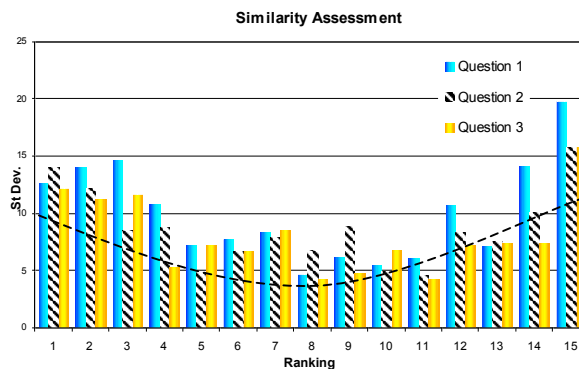


Figure 5. Standard deviations of rankings for Questions 1, 2 and 3.

Responses to Question 3 follow this pattern to a lesser extent and have a lower standard deviation corresponding to the lower coefficient of

concordance W . This is evident in the graph from the lower level of standard deviation for ranks 11-14.

4.1.2 Hypothesis 2: Assessment of similarity is context dependent

This hypothesis assumes that if a subject's judgment of similarity depends on context; their answers should vary across different design contexts. To test the hypothesis subjects' responses to all three questions are compared and since the same subjects answered the three questions, it is assumed that his or her responses need not be in 100% agreement in order to accept this hypothesis. The normalised responses were averaged and this average was compared for each ranked diagram across different contexts using Kendall's coefficient of concordance W . The value of W for Questions 1-3 is 0.18, which is low and suggests that rankings are not associated with a probability greater than 0.92.

To make sure that the similarity among contexts could not affect the result of the test statistic, subjects' responses for only Questions 2 and 3 are compared since context is defined more explicitly in these questions, i.e., in the context of a target diagram in Question 2 and in the context of a design task in Question 3. To compare rankings, the Spearman rank correlation coefficient (Gibbons 1976) was used as the test statistic. This test statistic is also a measure of association. As such, r_s should be equal to +1 when there is a perfect direct relationship between rankings. The value of r_s for Questions 2 and 3 is 0.14. Like the value of W , the value of r_s is low, which suggests that ranks are not associated. Such a low level of agreement suggests that subjects would not give the same evaluation under different contexts and confirms the hypothesis.

4.1.3 Hypothesis 3: The visuospatial information used to assess similarity will vary depending on context.

As the previous test demonstrated similarity assessments are dependent on context. This hypothesis assumes that if a subject's judgment of similarity depends on context, the visuospatial attributes used to assess similarity will also vary across different contexts. Subjects' responses to the types of visuospatial information were normalized and plotted in the graph shown in Figure 6.

The graph shows the normalised responses for each question and provides an indication of whether assessments are more associated with particular attributes. Looking at the graph there is no one single attribute category that can be seen to be consistently used by subjects across the three ranking tasks. Significantly combinations of attributes were most commonly described by subjects across all questions. This category had the highest average response with 30.5% of subjects listing multiple features. This was followed by subjects reporting that they could not describe any attribute/s,

with 22.5% specifying ‘unknown’. On average individual shape features scored the lowest number of responses with an average of 0.5% of responses describing some individual feature.

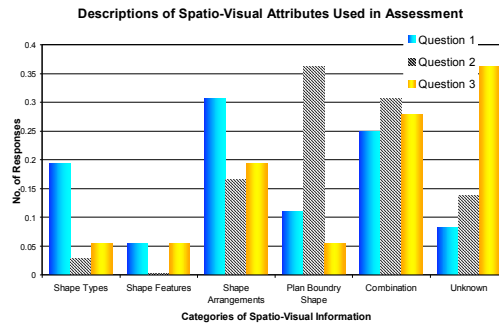


Figure 6. Responses of visuospatial attributes used in Questions 1, 2 and 3.

4.1.4 Hypothesis 4: An observer’s judgment of similarity should show some correlation with the results of the SOM networks.

This test compares subjects’ responses with the results of the Q-SOM model. The test assumes that the similarity and therefore types of features identified by the model may also be important in designer’s judgments. As the previous two tests demonstrated similarity assessments are dependent on context and the features used by subjects will also vary accordingly.

Comparisons are based on four networks, where two of the best performing networks and two of the worst performing networks were analysed. These networks were chosen based on the results from Precision Matrix, JAC and FM analysis methods. The hypothesis is tested with the Spearman rank correlation coefficient. Analysis is based on the normalized subject responses and compared to the ‘ranking’ derived from networks using the activation values of each SOM’s output nodes. The test statistic r_s for Questions 1 to 3 are shown in Table 3.

TABLE 3: Spearman rank correlation coefficients between subjects’ responses and the Q-SOM model with the two best performing networks.

PERFORMANCE	NETWORK	SPEARMANS RANK	Q 1	Q 2	Q 3
BEST	5X5 CFS with TF*IDF	$r_{s,n}$	0.62	0.70	0.38
	5X5 FB with TF*IDF	$r_{s,n}$	0.59	0.73	0.32
WORST	3X3 CFS with FREQ.	$r_{s,n}$	0.13	0.18	0.13
	3X3 FB with FREQ.	$r_{s,n}$	0.15	0.12	0.09

Table 3 shows all values of r_s vary significantly across all questions and networks. Values of r_s for Questions 1 and 2 and the two best performing

networks are all above 0.59, with the highest value of r_s for Question 2 with 0.73. These high values support the hypothesis that the subjects' responses and the Q-SOM model are associated with a probability of 0.78. However, the network which returned on average the highest values of correlation with subject responses was not the best performing network as determined by internal performance measures. Instead, the second best performing network, i.e., the 5x5 CFS network with TD*IDF weights has on average the highest values across all questions, with the exception of Question 2. As expected the values of r_s for Questions 1, 2 and 3 across all of the worst performing networks are significantly lower than those values of the best performing networks.

This hypothesis must however be rejected for Question 3 which uses contextually dependent similarity assessment. Here, values for the two best and worst performing networks are significantly lower with 0.38 in the best case and 0.09 in the worst.

4.1.5 Hypothesis 5: Correlations between an observer's judgment and computational assessment improves when context is not considered

Based on the confirmation of Hypothesis 4 for Question 1 and 2, this test uses the same approach to evaluation, i.e., the Spearman rank correlation coefficient, but considers only Question 1 which has no explicit contextual information. Recalling that Question 1 asks subjects to rank the similarity of diagrams based on their complexity, this test compares the subjects' responses and the results of the two best performing networks. The values of r_s are relatively high, which means that there is an association between the subjects' answers and the 5x5 CFS network with TD*IDF weights and 5x5 FB network with TD*IDF weights as shown in Table 3.

These results support the hypotheses that when context is not considered the two best performing networks are able to match designers' judgments. It is inconclusive as to whether networks can better match them when certain types of contexts are not considered. If we consider the value of r_s for Question 2, where context is implicitly defined in the target diagram, the correspondence is higher resulting in a greater association between subjects' responses and the results of the two best performing networks. If we consider the correlation between responses to Question 3 (a contextually dependent task) and the networks, it is significantly lower. The testing of this hypothesis is therefore inconclusive.

5. Discussion

The objective of these cognitive experiments was to evaluate correlations between the subjects themselves and between subjects and the results of the Q-SOM model. An important observation from the results reported here is that although subjects' responses are associated, in comparison to previous

experiments of feature-based similarity (e.g., 0.90 in Tversky's 1977 experiment) the degree of concordance (Section 4.1.1) among subjects' answers is only satisfactory, ranging from 0.21 in the worst case (Question 3) to 0.75 in the best case (Question 2). This lower degree of concordance may be due to the complexity and number of design diagrams that were evaluated and the use of a target diagram (Question 2) and design task (Question 3).

In other research on semantic similarity assessment human-subject experiments have been used to determine the effectiveness of computational models. These experiments found a correlation of 0.79 using an information content approach (Resnik 1999), and 0.83 using a distance approach (Jiang and Conrath 1997). The results reported here are not directly comparable, since here similarity is visuospatial, whilst these other studies evaluate similarity among the semantic relations of textual terms. Our results appear to support the use of the Q-SOM model for feature-based similarity of 2D diagrams, where correlation between the model and the subjects' answers was 0.32 in the worst case (contextually dependent) and 0.73 in the best case. Analysis shows the performance decreases when a design task is explicitly defined.

In cognitive studies of similarity assessment outside the design domain, it has been shown that goal-oriented similarity assessments vary with the task at hand (Landry et al. 2001). It might be concluded that since design scenarios are by nature 'open', ambiguous and interpretive that they result in other criteria, perhaps non-visual such as a designer's preferences and existing design experience, playing a larger role in assessment.

Other studies have shown that with practice, individuals develop representations of features that are useful for the task at hand and are treated as single units through processes of differentiation and unitization (Goldstone 2003). If a feature varies independently of others, or occurs more frequently, observers may develop a specialised detector for that feature (Goldstone, 2003). When specific information about a feature is known, it has been shown to take precedence (Heit and Rubinstein, 1994). In other studies of unsupervised category learning of visual patterns, experiments show that individuals are also sensitive to the frequency with which features of visual stimuli co-occur (Edelman et al 2001).

The results presented here have shown that during assessment, and especially in the context of a design task, visual sorting largely depends on combinations of visuospatial information and designers may not be able to describe the features used during assessment. Subjects were not able to describe the basis for their decision procedures. On average 22% of subjects could not specify the visual attributes used to assess similarity. This outcome could be expected in the design domain since the similarities which can be distinguished in and between design diagrams can be based on a variety of

attributes belonging to relatively complex geometries. Goldstone and Son (2005) have also illustrated that individuals are still able to make similarity assessments even when the exact properties of relevance are unknown.

Cognitive studies undertaken in design on the types of information categories utilised during drawing (Suwa and Tversky 1996, 1997) have shown that designers process information at multiple levels and shift focus between varieties of information categories based on their current strategy. Using a retrospective protocol analysis Suwa and Tversky (1997) found that architects are able to perceive visual attributes in diagrams such as shapes and angles and are able to “read-off” more abstract features and responses, pursuing design thoughts more deeply. They concluded that the design process consists of cycles of focus shift, supporting the view of a two-way bottom up and top down approach to perception. This supports the multi-level approach to information processing adopted by the Q-SOM model where both local and global visuospatial information is utilised.

The results obtained from attribute descriptions made by subjects also suggest that the local-versus-global argument is ill-posed and that either one may dominate depending on the intention of the designer. According to Kinchla et al (1983), the observer has the ability to select alternative visual strategies, sometimes attending to the local details and sometimes attending to the global properties. The multiple levels of abstraction that are supported by the Q-SOM model enable both low-level geometric and high-level semantic feature to be utilised by the network for similarity assessment.

Finally, although our results found no significant differences in how 1st year versus final year design students assessed the similarity of Wright’s diagrams, it was informally noted that the majority of 1st year design students finished prior to the time allotted whilst almost all final year designer students required the full time allocated to complete each experiment. Other researchers have also shown that knowledge and design experience can influence the comprehension of graphic information (Gobert 1999) as well as the type and frequency of those visual features detected (Akin 1978, 1986).

6. Conclusion

This paper has investigated how human-subjects (designers) assess the similarity of 2D design diagrams using a variety of assessment techniques. The overall objective of this research is to increase the Q-SOM model’s performance through further cognitive evaluations and computational experimentation so as to find those network and feature variables, which on average, perform as well as possible in relation cognitive assessments.

Following recommendations made by Cook and Campbell (1979) on the types of validity which may be attained for models of causality, it is

proposed that analogous dimensions of validation be followed here. If it can be shown that the Q-SOM model embodies all of the following four dimensions of validation then the approach is more likely to be of utility and significance. Two of the four dimensions have already been tested here and based on our results can be explored further, namely:

- (i) *statistical conclusion validity*, i.e., are there similarities between the diagrams?
- (ii) *external validity*, i.e., given that there are similarities, how generalisable are these results across different network types, persons and contexts?

In addition, two other important dimensions must now be tested, including:

- (iii) *internal validity* i.e., given that there are similarities, is it causal from particular operational variables?
- (iv) *construct validity* i.e., given that similarity is causal, what are the particular cause and effect constructs involved?

Utilising these dimensions of validation will ensure that the Q-SOM model will be capable of measuring similarities of 2D design diagrams in a cognitively congruent manner. This will also allow explicit examination of how particular network, featural and contextual variables affect similarity assessments. Patterns of information flow within the network can then be examined, allowing for hypotheses concerning the type and weights of features used to distinguish 2D diagrams to be investigated in relation to those features described by designers as being salient in assessments.

References

- Akin, Ö: 1986, *The Psychology of Architectural Design*, Pion, London.
- Akin, Ö: 1978, How do architects design? in J.-C. Latombe (ed.), *Artificial Intelligence and Pattern Recognition in Computer-Aided Design*, North-Holland, New York, pp. 65-104.
- Cook, T, and Campbell, D: 1979, *Quasi-experimentation: Design and Analysis Issues for Field Settings*, Boston, Houghton Mifflin.
- Daniel, W: 1978, *Applied Nonparametric Statistics*, Houghton and Mifflin, Boston MA.
- Downton, M and Brennan, T: 1980, Comparing classifications: an evaluation of several coefficient of partition agreement, *Proceedings Meeting of the Classification Society*, Boulder, CO. pp 418-425.
- Edelman, S, Hiles, BP, Yang, H and Intrator, N: 2001, Probabilistic principles in unsupervised learning of visual structure: human data and a model, in S Becker (ed.) *Proceedings of the 2001 Conference on Neural Information Processing Systems (NIPS)*, MIT Press, Cambridge, MA.
- Fowlkes, E and Mallows, C: 1983, A method for comparing two hierarchical clusterings, *Journal of American Statistical Association* **78**: 553-569.
- Gobert, JD: 1999, Expertise in the comprehension of architectural plans: Knowledge acquisition and inference making, in J Gero and B Tversky, (eds.) *Visual and Spatial Reasoning in Design*, Sydney, Australia, Key Centre of Design Computing and Cognition, pp. 185-205.

- Goldstone, R: 1994, Similarity, interactive activation and mapping, *Journal of Experimental Psychology: Learning Memory and Cognition*, **20** (1): 3-28.
- Goldstone, R: 2003, Learning to perceive while perceiving to learn, in R Kimchi, M Behrmann and C Olson (eds.) *Perceptual organization in vision: Behavioral and neural perspectives*, Lawrence Erlbaum, Mahwah, NJ, pp. 233-278.
- Goldstone, RL and Son, JY: 2005, Similarity, in KJ Holyoak and R Morrison (eds.) *Cambridge Handbook of thinking and reasoning*, Cambridge University Press, Cambridge, UK.
- Heit, E and Rubinstein, J: 1994, Similarity and property effects in inductive reasoning, *Journal of Experimental Psychology: Learning, Memory, and Cognition* **20**: 411-422.
- Jiang, J and Conrath, D: 1997, Similarity based on corpus statistics and lexical taxonomy, *International Conference on Computational Linguistics (ROCLING X)* Taiwan, pp 19-35.
- Jupp, J and Gero, JS: (submitted), Visual style: qualitative and context dependent categorisation, *Artificial Intelligence for Engineering Design, Analysis and Manufacturing AIEDAM* Special Issue Understanding, Representing and Reasoning about Style.
- Jupp, J: 2005, *Diagrammatic Reasoning in Design: Computational and Cognitive Studies in Similarity Assessment*, unpublished PhD thesis, Key Centre of Design Computing and Cognition, University of Sydney, Australia.
- Kinchla, RA, Solis-Machias, V and Hoffman, J: 1983, Attending to different levels of structure in a visual image, *Perception and Psychophysics*, **33**, 1-10.
- Lamberts, K and Shanks, DR, (eds): 1997, Knowledge, concepts, and categories, *Psychology Press*, Hove, UK.
- Landry, SJ, Sheridan, TB and Yufik, YM: 2001, A methodology for studying cognitive groupings in a target-tracking tasks, *IEEE Transactions on Intelligent Transportation Systems* **2**(2): 92-100.
- Medin, DL, Goldstone, RL, and Gentner, D: 1993, Respects for similarity, *Psychological Review*, **100**: 254-278.
- Medin, DL, Lynch, EB and Solomon, KO: 2000, Are there kinds of concepts? *Annual Review of Psychology* **51**: 121-147.
- Palmeri, TJ and Gauthier, I: 2004, Visual object understanding, *Nature Reviews Neuroscience* **5**, 291-304.
- Meeran, S, and Pratt, MJ: 1993 Automated feature recognition from 2D drawings, *CAD*, **25**(1): 7-17.
- Resnick, O: 1999, Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity and natural language, *Journal of Artificial Intelligence Research*, **11**, 95-130.
- Slonim, N, Friedman, N and Tishby, N: 2002, Unsupervised document classification using sequential information maximization, *Proceedings of SIGIR '02, 25th ACM international Conference on Research and Development of Information Retrieval*, Tampere, Finland, ACM Press, New York, USA.
- Suwa, M and Tversky, B: 1996, What architects see in their sketches: Implications for design tools, *Proceedings of CHI '96*, pp. 191-192.
- Suwa, M and Tversky, B: 1997, What architects and students perceive in their sketches: A protocol analysis, *Design Studies*, **18**: 385-403.
- Thomas, MSC and Mareschal, D: 1997, Connectionism and psychological notions of similarity, *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*, London: Erlbaum, pp. 757-762.
- Tversky, A: 1977, Features of similarity, *Psychological Review* **84**: 327-352.