Using Semantic Similarity to Predict Angle and Distance of Objects in Images

Sterling Somers
Institute of Cognitive Science, Carleton University
1125 Colonel By Drive, Ottawa, ON K1S 6B6
Canada
sterling@sterlingsomers.com

Jonathan Gagné
Dept. Systems Design Engineering
University of Waterloo
200 University Ave West
Waterloo, ON N2L 3G1
Canada
jgagne@engmail.uwaterloo.ca

Cesar Astudillo
Universidad de Talca
Facultad de Ingeniería
Departamento de Ciencias de la Computación, Km. 1
Camino a los Niches, Curico, Chile
castudillo@utalca.cl

Jim Davies
Institute of Cognitive Science, Carleton University
1125 Colonel By Drive, Ottawa, ON K1S 6B6
Canada
jim@jimdavies.org

ABSTRACT
A presentation of an Artificial Intelligence (AI) called Visuo that stores and guesses quantitative visual-spatial magnitudes (e.g., sizes of objects). In this analysis, Visuo is used to store polar (angle and distance) relationships between objects in images. It uses a database of tagged images as its memory and approximates unexperienced magnitudes by analogy with semantically related concepts. This shows the transferring of information from high semantically related concepts yielding significantly higher accuracy in angle and distance estimations over using medium or low semantically similar items.

Author Keywords
Creativity, imagination, visualization, analogy.

ACM Classification Keywords
I.2.0 General---Cognitive simulation

I.6.4 Model Validation and Analysis.

General Terms
Theory.

INTRODUCTION
Human visual imagination is an incredibly complex cognitive phenomenon. Imagining a scene draws on our memory to populate the scene with relevant scene objects in varying fidelity. Recent evidence suggests further that the same brain regions involved in vision are also involved in visual imagination, suggesting that we in fact, in some sense, see the products of our own mind [5, 7]. When imagining a scene intended to mimic the real world, the creative process has to produce objects of an appropriate colour, size, *et cetera* to fit the intended context of the scene. Imagined scenes are populated with the right kinds of objects for the intended context. Furthermore, how objects in an imagined scene relate to each other spatially is typically determined realistically. For example, when imagining a bird in the sky, such a scene might also involve secondary objects such as trees, the ground, the sun, among others. All elements included in the scene will have spatial relationships to each other. In an image of such a scene, the bird in the sky will be an appropriate angle and distance from the trees, while the trees will be an appropriate angle and distance from the sky and the ground, and on it continues until all objects are spatially related to each other.

Our theory is that visual imagination exploits regularities in experience. When we imagine a realistic scene, the visual-spatial properties of that scene end up being the way they seem because that is how they occur to us through our experience. Grass is on the ground, trees are above the ground, and the sky is above the trees; and that is how we experience them. When we take photographs, these regularities get stored digitially (see Figure 1). We propose that an AI designed to perform human-like imagination can exploit the visual-spatial information contained in digital photographs to inform the visualization of imagined scenes.

There are times, however, when we might want to imagine a scene containing objects in contexts we have never experienced them in. Extrapolating from the example above, perhaps the imagined bird in the scene is a raven. Although you have never seen one outside of a bird identification guide, you know, more or less, what a raven looks like. You might know that crows, which are often seen, are very similar to ravens. Could an analogy relating ravens to crows be used to fill in the details of our imagined scene with the raven? Apart from looking similar, crows are likely to have similar spatial relationships to trees and the other elements of the scene. Would the same likely hold true if the analogous object was less similar? Would any bird have similar spatial relationships? Would any animal?

C&C'11, November 3–6, 2011, Atlanta, Georgia, USA.

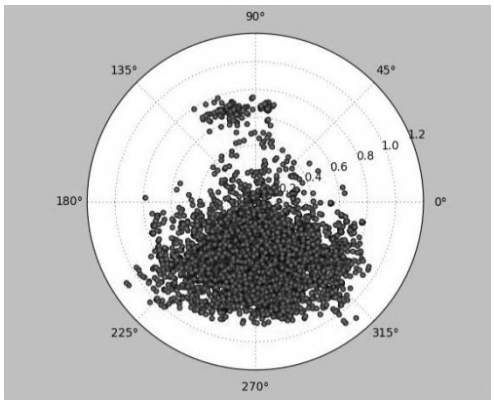

Figure 1. 'tree-grass'. To interpret this picture, consider the centre to represent 'tree'. Each dot represents an instance of the relationship between 'tree' and 'grass' in our image database. As expected, the grass appears below trees in most cases.

This paper presents an extension to a Python program, Visuo, intended to model visual-spatial instantiation [2,5]. Shown already to predict sizes of analogous objects when using size modifying adjectives (e.g., 'large'), for the present work, Visuo has been extended to store angles and distances. This paper will provide an answer to the above questions by evaluating Visuo's storage of angle and distance relationships between two objects of high, medium, and low semantic similarity. We hypothesize that the spatial relationships between items of high semantic similarity will approximate a target relationship with greater accuracy than those of medium or low semantic similarity.

VISUO

Visuo is a program written in Python intended to be a model of memory for quantitative data, and as a model for re-instantiating quantitative information from known priors as well as using analogy to instantiate quantitative information for cases that have not been experienced. For example, Visuo has been shown to predict the size of target objects, such as *large raven*, when it had information on *raven* sizes but not the target *large raven*. Having data on the size of one object, say all ravens, data on a semantically similar object, say the size of crows, and data on the size of the same semantically similar object with a linguistic modifier, such as *large crow,* Visuo was able to predict the size of a *large raven* [2].

Visuo implements two phases: a training phase and a visualization phase. While previous papers have addressed both phases, this paper is primarily concerned with the training phase and comparing the results thereof. While we have developed a method for instantiating spatial relationships, this method remains largely untested. Therefore, our discussion of theory will be limited to the assumptions and processes of the training phase.

Visuo is trained by reading text files with numerous entries. In this particular application of Visuo, the text file describes spatial relationship data (angle and distance) between object pairs. For example, a single entry could be for the spatial relationship of a raven to a tree, consisting of a description, "raven <relationship> tree"; a value for angle, "angle=27.45"; and a value for the distance between the two items, "distance=200.43". Each entry read by Visuo is called an 'experience,' Angles are measured in degrees and distances are percentages of a picture divided by 100. For example, the distance of 200.43 is approximately 2% of an image away. All angles and distances used to train Visuo were taken from images of uniform size.

Training

Visuo implements two types of memory: episodic memory and semantic memory. Exemplars, Visuo's implementation of episodic memories, represent memories of objects occurring at a specific place and time [12]. As Visuo's episodic memory is not involved in the present analysis, we will concentrate on its semantic memory.

Semantic Memories are memories of general concepts abstracted from specific instances [11]. When Visuo experiences a relationship for the first time both an episodic and a semantic memory are formed. With every following instance of the same general relationship, new episodic memories are formed, and the semantic memory is modified, with the new information being incorporated with the old. Semantic memories in Visuo are called prototypes.

Distribution of Fuzzy Set Membership

While Visuo experiences precise quantitative values, the precision is not stored in semantic memory. Behavioural data show that people represent with graded membership in categories [9]. To account for this, each perceptual detector in Visuo represents using fuzzy-set perceptual categories. In fuzzy set theory, the membership of an instance in a given set is described with a fuzzy membership value ranging from 0 (clearly not in the set) to 1 (clearly a member of the set). In the present analysis, Visuo employs a category set for distance and a category set for angle. For a relevant example, a *crisp* input of "10" becomes a vector of membership values in for a set of *fuzzy* number categories [4]. We will describe what exactly these numbers are in the following subsections.

Every crisp number gets distributed as a member of all fuzzy number categories to *some degree*. Membership degrees are represented by a real number between 0 and 1. For example, a crisp number 10 would have a high membership in a fuzzy number set 8, and a lower membership in a fuzzy number sets for 2 and 15. *Fuzzification* is the process of transforming a crisp number to a fuzzy number. A separate distribution is created for each distance and angle pair.

Representation of Distance

Dehaene, et al. [3] provides evidence that people naturally (without educational intervention) represent distance

logarithmically. Therefore, Visuo uses a 15 point, approximate logarithmic scale (0, 2, 5, 10, 20, 35, 65, 100, 160, 150, 400, 600, 900, 1350, 1800) as categories for distances. The fuzzy categories have overlapping ranges and input values will be distributed across relevant categories. For example, a distance of 10 (1% of a picture in distance) would be represented as the vector (0.0, 0.0, 0.67, 1.00, 0.33, 0.0, …). Each value in that vector represents a membership in the fuzzy number set described above.

Representation of Angle

Huttenlocher, Hedges, and Duncan [10] provide evidence for the use of polar coordinates (angle and distance) as well as angular categories in object location estimation. Visuo uses a 16-point scale with 22.5-degree intervals for estimating angle (180, 157.5, 135, 112.5, 90, 67.5, 45, 22.5, 0, -22.5, -45, -67.5, -90, -112.5, -135, -157.5). Similarly, the fuzzy angular categories have overlapping ranges, with values distributed across relevant categories. For example, an angle of 45 degrees is represented by the vector (0, 0, 0, 0, 0, 0.5, 1.0, 0.5, 0, …). Each value represents degree of membership in the fuzzy number sets (e.g., 0.5 membership in 22.5-degrees set).

Incorporating New Data

Visuo creates a prototype for each tag pair (e.g., tree-grass). For each novel prototype experienced, Visuo creates a fuzzy distribution for distance and a fuzzy distribution for angle. Instead of creating a new vector for angle and distance, for each repeated tag pair, Visuo incorporates the new data into the old. When a new example of a previously experience tag pair is observed, each fuzzy membership number in the distribution is replaced by a recursive average:

$$v_{avg} = \frac{(n * v_{old}) + v_{new}}{n + 1}$$

where v_{old} is the previous value, v_{new} is the new value, and n is the number of experiences. Following the calculation of v_{avg}, the count n is increased by one. For each fuzzy number in the vector, the prototype represents the mean value of the membership for exemplars for the corresponding fuzzy number. In this way, the prototype represents and average of all experiences. For example, there would be a prototype storing a distribution representing all of the angles ever experienced between a tree and grass.

Inferring Missing Data Through Analogy

While Visuo uses prototypes to instantiate visual descriptions of scenes, it need not rely on exact matches in memory of the intended visualization. When data is missing from Visuo's experience, it will use analogy (in this case, semantic similarity) to approximate the desired data. For example, in the database used in the present analysis (see below), there is no instance of a raven above grass in Visuo's experience. If Visuo is asked to visualize a raven above grass, instead of reporting that no data is available, it will find the closest semantically related item to one of the targets and use that data as an approximation. Visuo uses the Wu-Palmer similarity measure [14] as implemented in NLTK version [1] of WordNet [8]. The Wu-Palmer similarity measure is squared to help bias for similarity.

EVALUATION

Data

The Peekaboom [13] image database is a database with point clouds and tags for elements in the images. We used approximately 50,000 images from the Peekaboom database and mined approximately 200,000 unique spatial relationships, for both angle and distance between tagged item pairs. We used the centroids of the point clouds as an indicator of the location of each tagged element. The database uses images from the internet, including photographs, advertisements, drawings, etc. It was not filtered.

Of the ~200,000 unique spatial relationships (tag pairs), we chose to work with the 100 pairs with the largest number of instances to maximize the number of data points per tag pair. These top 100 spatial relationships are called the target pairs. Of each target pair, the first tag is the static tag and the second the dynamic tag. To determine semantic relatedness, the static tag remained the same and the dynamic tag was replaced. For example, with the pair *grass-raven*, 'grass' would be the static element while a suitable match, based on semantic similarity, would be found for 'raven'. Using the data from our database, 'crow' has the highest similarity to raven and the prototype of *grass-crow* would be used as the source concept for the analogy.

After selecting the target pairs, three test replacements for the dynamic targets were found and sorted into groups of **high** (Wu-Palmer similarity^2 0.7 – 0.9), **medium** (Wu-Palmer similarity^2 0.34 – 0.69), and **low** (Wu-Palmer similarity^2 0 – 0.33) similarity. The new targets were selected to maximize the degree to which they represented their group: maximizing similarity for the high group, maximizing closeness to the middle for the medium group, and as dissimilar as possible for the low group. A minimum of 25 data points was required for replacement tags to be included.

	Mean Membership Difference, High Semantic Similarity	Mean Membership Difference, Medium Semantic Similarity	Mean Membership Difference, Low Semantic Similarity
Distance Membership	0.0377	0.0685	0.0971
Angle Membership	0.0634	0.0924	0.1159

Table 1. Mean fuzzy membership difference for high, medium, and low semantic similarity groups.

Procedure

For each of the 100 target pairs, Visuo was trained on all the pairs and the vectors for angle and distance for each pair were stored in text files. The training was conducted across two different computers, utilizing 2 CPUs on each computer. Due to a file-reading error resultant from the attempted concurrent use of a single database file, 1 target pair was excluded from the present analysis. Thus, a total of 99 target pairs, resulted in 4 (target, high, medium, and low) × 99 = 396 training sessions. Each training session produced one distribution for angle and one distribution for distance, resulting in 792 distributions.

If our hypothesis is correct, the predicted angles and distances between the elements of each target pair should be closest to the actual angles and distances when analogizing from the high-similarity group, and furthest when analogizing from the low-similarity group. The medium-similarity group's accuracy should fall somewhere in between.

Results

Angle

The high-angle group was a significantly better approximation than the medium-angle group (Wilcoxon test, z = -10.77, p < .01, r = -1.08). The low-angle group was also significantly worse than the medium-angle group (z = -8.34, p < .01, r = -0.84). The high-angle group was also significantly better than the low-angle group (z = -18.62, p < .01, r = -1.87).

For statistical analysis, data was grouped on a per-test-pair basis. For each test pair, categories for distance were coded 1-15, allowing us to group each logarithmic distance category for each test pair. Likewise, categories for angle were coded 1-16, categorizing the angle categories. Because the present analysis is aimed at testing how well pairs in the different groups approximate the target pairs, differences between the target and high, target and medium, and target and low pairs were calculated. Figure 2 illustrates a comparison for target, high, medium, and low groups for one sample target pair. Table 1 contains a summary of the mean differences across membership vectors for all 99 item pairs. While the mean differences may seem small, recall that the 0-1 range is distributed across 15 items and 16 items for distance and angle, respectively. For perspective, the average membership for the target pair in the angle group is 0.125 and the average membership for the target pair in the distance group is 0.130. An average difference of 0.1159 (see Table 1) is larger than it might initially appear without this perspective. As shown in Table 1, the mean membership decreases (increasing difference) as semantic similarity decreases.

Distance

Wilcoxon tests between the high-distance group and the medium-distance group revealed that the high group was significantly more accurate than the medium group, (z = -10.73, p < .01, r = 1.08). The low-distance group was also significantly less accurate than the medium-distance group, (z = -7.43, p < .01, r = 0.75). The high-distance group was also significantly more accurate than the low-distance group (z = -15.96, p < .01, r = 1.67).

Discussion

The results above are very encouraging as a preliminary analysis into the use of analogy to supplement spatial relationship data. Moreover, they are particularly promising when one considers the sources of error and variance.

The first source of error comes directly from our data. Due to the nature of the Peekaboom data collection (point cloud creation), the point clouds around the elements in a picture are inexact. Furthermore, the outside boundary (convex

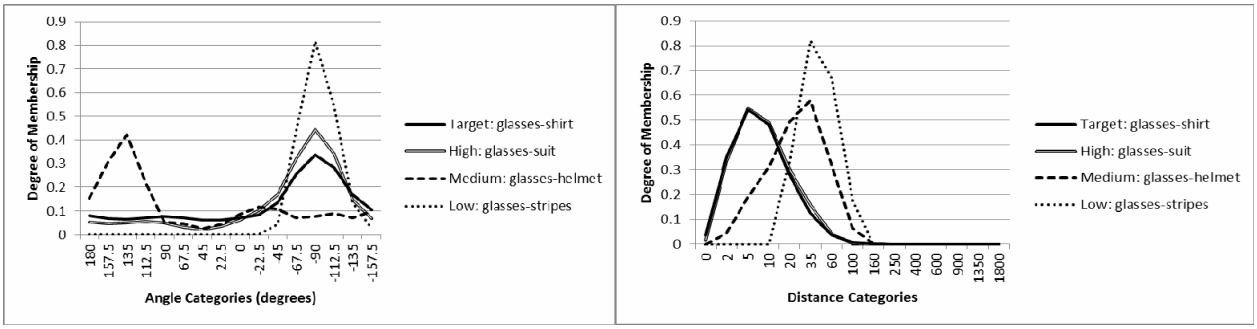

Figure 2. Angle (left) and distance (right) vector comparison for target pair *glasses-shirt*.

hull) of point cloud may not perfectly outline the tagged object it is meant to identify. These point clouds were used to calculate a centroid and therefore, the larger the error in the point cloud, the larger the error in the centroid.

Secondly, due to availability of data, there was a good deal of variance in the precise similarity values within a group. For example, there are instances where tag pairs that have a high degree of semantically similar are at the lower bound of the high category and medium semantically similar items are at the upper bound of the medium category, which results in items with similar semantic similarities to the dynamic target. Although our results suggest that this source of variance is acceptable in the present analysis, future work will address this issue to maximize semantic similarity. In this future analysis, we would expect a close approximation of the target group by items restricted to some minimum semantic similarity. That said, there are still a number of data filtering issues yet to resolve.

One issue yet to resolve is the use of centroids as specifying an object's location. Using centroids is a simple approach in determining an object's location but can be misleading in terms of data extraction. For instance, in an image where a bird appears in the sky, the centroids become meaningless as the sky surrounds the bird. More complex relationships such as 'in' will require specific detectors that go beyond angle and distance.

CONCLUSION

Human visual imagination is an important and largely unaddressed problem in the cognitive sciences. While creating computer programs that can combine multiple experiences to produce imagination-like scenes presents an intriguing software engineering problem space, creating programs that can create scenes in novel and realistic ways goes beyond that and begins to explore the nature of creativity. Our running theme throughout this paper has been to advocate the use of analogy to apply semantically related data to visualizations otherwise impossible for an AI to instantiate. While one AI approach to our problem space is to get larger and larger databases, we aim to adopt more human-like solutions.

We have shown in this paper that using high semantic similarity is a plausible candidate for analogical visual-spatial reasoning. By showing that high semantic similar items more closely approximate angle and distance data than do medium or low semantic similar items, we have given reason to investigate this use of analogy further. For instance, an appropriate next step would be to determine what minimal degree of similarity is required for analogous relationships to reliably approximate their targets. Furthermore, qualitative analyses on the distributions such as the ones in Figure 1 has suggested that there is perhaps some form of categorizing possible for the spatial relationships as some distributions seem ideal for visualization, while other distributions appear flat.

The work that has been presented here represents a module in a much larger project of simulating visual imagination. As discussed in the introduction, knowing how elements in a novel image should be organized spatially is an essential step in having AIs capable of producing realistic scenes. Also part of our aim is to explain how human imagination works. It is our hope that by implementing a model of imagination, that the model can act as a tool for guiding the scientific inquiry into the nature of human imagination.

Visuo, we expect, will be at the heart of our imagination model. Visuo itself represents a theory about how we learn certain concepts and relationships, and how we store and retrieve quantitative data. Visuo does not explicitly make rule-based concepts such as 'trees are always above the grass.' Instead, rules that capture the regularities of experience emerge as an average of experience. The cluster of dots in Figure 1, for example, provide compelling qualitative evidence that at least some concepts might emerge that way. That some method of storing and retrieving quantitative data is required for visual imagination is obvious, whether the theory behind Visuo holds is yet to be seen. While in the present study we have validated the approach of using semantic similarity to supplement a database, we are also carrying psychological experiments to determine if humans use similar approaches when approximating data.

REFERENCES

1. Bird, S. and Loper, E. NLTK: The Natural Language Toolkit. *Proceedings of the ACL demonstration session*, (Barcelona, Spain, 2004), Association for Computational Linguistics, 214-217.
2. Davies, J. and Gagné, J. Estimating quantitative magnitudes using semantic similarity. *Twenty-Fourth Conference of the American Association for Artificial Intelligence workshop on Visual Representations and Reasoning* (Atlanta, GA, 2010), 14-19.
3. Dehaene, S., Izard, V., Spelke, E., and Pica, P. Log or Linear? Distinct Intuitions of the Number Scale in Western and Amazonian Indigenous Cultures. *Science, 320* (5880) 1217-1220.
4. Dubois, D. and Prade, H. Fuzzy numbers: an overview. in J.C. Bezdek *Analysis of Fuzzy Information Vol. I: Mathematics and Logic*, CRC Press Boca Raton, 1987, 3-39.
5. Gagné, J. and Davies, J. Visuo: A model of visuospatial instantiation of quantitative magnitudes. *Knowledge Engineering Review. Special Issue on Visual Reasoning*, forthcoming.
6. Kosslyn, S. M. Image and Brain: The Resolution of the Imagery Debate. MIT Press, Cambridge, 1994.
7. Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M., Morito, Y., Tanabe, H. C., Norihiro, S. and Kamatani, Y. Visual image reconstruction from human brain

activity: A modular decoding approach. *Neuron*, *60* (5), 915-929.

8. Fellbaum, C. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, 1998.
9. Hampton, J. A. Typicality, Graded Membership, and Vagueness. *Cognitive Science*, *31* (3), 355-384
10. Huttenlocher, J., Hedges, L. V., and Duncan, S. Categories and Particulars: Prototype Effects in Estimating Spatial Location. *Psychology Review, 98* (3), 352-376.
11. Rosch, E.H. Natural categories. *Cognitive Psychology*, *4* (3), 328-350.
12. Tulving, E. Précis of Elements of Episodic Memory. *Behavioral and Brain Sciences*, *7*, 223–268
13. Von Ahn, L., Liu, R. and Blum, M. Peekaboom: A Game for Locating Objects In Images. *Computer-Human Interaction Conference*, (Montréal, Canada, 2006) 55-64.
14. Wu, Z. and Palmer, M. Verb semantics and lexical selection. *32nd Annual Meeting of the Association for Computational Linguistics*, (Las Cruces, New Mexico, 1994), 133-138.