How different are the visual representations used for object recognition in middle childhood and adulthood?

Dean Petters[1], John Hummel[2], Martin Jüttner[3], Ellie Wakui[4], Jules Davidoff[5]

Abstract. Recent experimental studies have shown that development towards adult performance levels in configural processing in object recognition is delayed through middle childhood. Whilst part-changes to animal and artefact stimuli are processed with similar to adult levels of accuracy from 7 years of age, relative size changes to stimuli result in a significant decrease in relative performance for participants aged between 7 and 10. Two sets of computational experiments were run using the JIM3 artificial neural network with adult and 'immature' versions to simulate these results. One set progressively decreased the number of neurons involved in the representation of view-independent metric relations within multi-geon objects. A second set of computational experiments involved decreasing the number of neurons that represent view-dependent (non-relational) object attributes in JIM3's Surface Map. The simulation results which show the best qualitative match to empirical data occurred when artificial neurons representing metric-precision relations were entirely eliminated. These results therefore provide further evidence for the late development of relational processing in object recognition and suggest that children in middle childhood may recognise objects without forming structural description representations.

1 Introduction

Compositionality is the property that the meaning of any linguistic or logical expression with multiple parts is determined not just by the meanings of those parts but the way they are put together. In addition to language, compositionality is found is a diverse range of other entities in the world. In our interactions with objects the perception of compositionality can be manifested across multiple modalities [14]. We can perceive visual compositionality in scenes and objects and thus form structural descriptions. So objects can be recognised by relations between their components parts [2].

The 'Recognition by Components' (RBC) theory of object recognition is distinguished from other structural description theories of object recognition because it postulates that geons (geometric components derived from readily detectable properties of edges) are the fundamental unit of representation in objects [2]. Geons can therefore be compared to phonemes in spoken language. In both systems, a small number of representational primitives can code for a very large number of component representations (words or visual objects, respectively). In the original RBC theory 36 geons are proposed as components for all objects, compared with the 55 phonemes required to represent virtually all words in human speech [2]. A key similarity of these systems is that how the primitives are combined matters. One way in which phonemes and geons differ is that phonemes form words by linkage in serial chains where the order matters. However, visual objects can be formed of multiple geons with several different types of relations, such as larger-smaller, and above-below or beside.

Artificial Neural Networks can represent visual compositionality and hence model natural cognition [8, 15]. Visual compositionality is also of interest in machine representations because it can facilitate artificial systems extracting verbal descriptions of scenes or objects. Active research questions include the comparative benefits of mechanisms for neural instantiation of visual combinatorial representations [15, 4], and how generalised shape information develops [5].

2 Recent empirical results in the development of configural processing in object recognition

A number of behavioural studies suggest there is a retarded developmental trajectory for object recognition - with object recognition skills continuing to significantly improve during adolescence [3, 14, 11]. Recently, Jüttner *et al* [12] examined developmental trends associated with identification of correct pictures when presented alongside incorrect distracters (in a 3 AFC task). Two distracter types were compared: part-changed stimuli, where one part of the stimuli was substituted for an incorrect part (figure 1); and a change to the overall proportions of the object (the configural change condition, figure 2).

Figure 1. Showing an animal version of a part change stimuli used in human studies. Selecting the 'real' cow image is a non-configural task as only one object-part needs to be checked at a time.

Figure 2. Showing an animal version of a relative size change stimuli used in human studies. Selecting the 'real' fly image is a configural task as recognition results from checking the relative sizes of two (or more) parts.

[1] University of Northampton, UK, email: dean.petters@northampton.ac.uk
[2] University of Illinios at Urbana-Champagne, USA
[3] University of Aston, UK
[4] University of East London, UK
[5] Goldsmith's College, London, UK

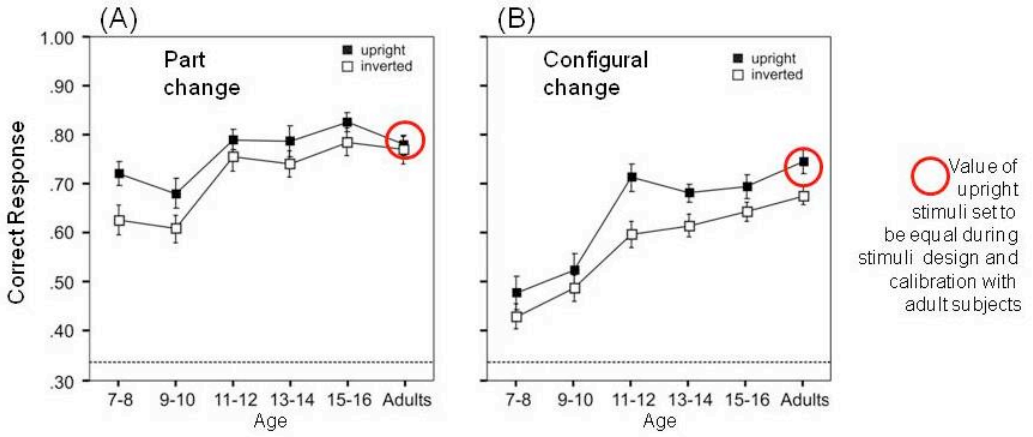

Figure 3. Results of experiments where participants of different ages were tested with part and configural changed stimuli.

In both part change and configural (relative size) change conditions, the task is to choose the 'correct' image. So in figure 1 the bottom 'cow' is the only image with a cows head. In figure 2 the middle 'fly' is the only one with eyes that are the correct size in proportion to its body. In addition to stimuli derived from a set of naturalistic animal images, experiments were undertaken with stimuli from naturalistic images of defined-base, rigid artefacts (see [12] page 163, for examples). Responses to defined-base, rigid artefact stimuli showed the same pattern of results to the animal stimuli.

The part change and configural change sets of experimental stimuli were calibrated to be equally difficult for adults - with an 0.8 mean accuracy set for both conditions. After calibration with adults on upright stimuli, adult performance was recorded on inverted (upside down) versions of the stimuli. Then the same stimuli set was used to assess recognition performance in school children aged between 7 - and 16- years of age in upright and inverted conditions. Overall, 32 participants were used in each of 6 age ranges (7-8, 9-10, 11-12, 13-14, 15-16, and adult).

The full description of method and results for these experiments is detailed in [12]. Performance in terms of accuracy, and latency preceding a correct response, show a similar pattern of results to each other, with no evidence of a speed/accuracy trade-off. The key empirical results for younger children (7-10 year olds) are that whilst part-change performance is marginally lower than adult levels, relative size change performance is significantly lower. For older children (11-16 year olds), part change performance has reached the adult level whilst relative size change performance is still not fully consolidate [12]. Figure 3 shows mean and standard errors of the recognition accuracy, with results combined across animals and artefacts as the stimulus type (animal/artefact) did not significantly affect recognition accuracy or latency nor interact with any other experimental variable. To evaluate a possible dual process explanation for these results this paper now presents simulation results gained by developmentally regressing JIM3 [8], a prominent dual process model that simulates visual object recognition.

3 Overview of JIM3

3.1 Introduction to JIM3

JIM3 is an eight layer Artificial Neural Network model of visual object recognition [9, 10, 8]. It takes as input a representation of contours from a single object's image. The output is a representation of an object's identity. Figure 4 (adapted from [8]) shows JIM3's 8 layers and the two places where changes were made to developmentally regress the architecture.

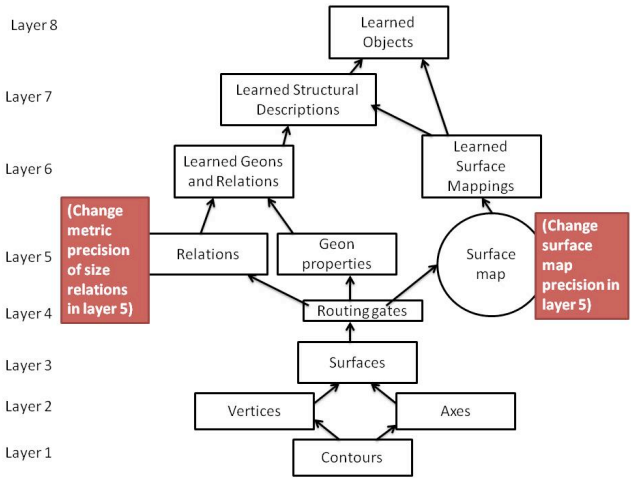

Figure 4. Diagram of JIM3 showing the two locations in the architecture where changes were made to capture this architecture's performance for an earlier developmental stage.

3.2 Layers 1 to 3 From feature maps to independent geons

The first three layers are comprised of feature maps and are concerned with grouping local features into sets. These sets correspond to which geons the features arise from. Layer 1 outputs the contours

present in the image. Layer 2 uses these contours to compute vertices and axes which are then processed by layer 3 as it computes the surfaces that belong to each geon. So the overall behaviour of this subsystem is to determine what individual geons are present in an image from the simultaneous presentation of a complete multi-geon contour set. These individual geons are then output from this subsystem as isolated and independent object parts with no explicit relationship to other geons arising from the same object.

When an object is initially presented to the model all the features of an image will tend to fire at once. This event is simulating the first tens of ms of natural object perception and occurs in the running simulation in the first several processing iterations. Then in an attentive process which involves inhibition and competition the attributes from different geons become temporally separated. This process occurs through the global action of a particular kind of artificial neural network connection termed by [9] as Fast Enabling Links (FELs).

The first three layers of JIM3 three act together to output each component geon at a different point in time. If this did not happen and attributes of separate geons fired synchronously then their attributes would get super-imposed. The three conditions which cause FELs to treat units as from the same geon are: local course coding of image contours; cotermination in a intra-geon vertex; and, distant collinearity through lone terminations. The simultaneously firing features become organised so that only the attributes for a single geon fire at one time by an iterative process of competition and inhibition.

3.3 Layer 4 Routing gates (Passing each independent geon forward separated in time from the other geons)

The 4th layer is a set of routing gates that splits the output from the first three layers and sends this output to two separate subsystems in layer 5. The information carried by these routing gates is of attribute sets for individual geons. After an initial period of phase locking the information about individual geons are sent as temporally separated signals. That is, geons attributes for individual geons are transmitted together and separated in time from the transmission of attributes describing the other geons present in the target object. This means that at any particular time the output from the routing gates is just an attribute set for one individual geon from the target object. Then after a gap in time the next geon is transmitted. Then after further gaps in time more geons are transmitted until the details of all geons present in the target object are communicated through these routing gates.

3.4 Layer 5 View dependent and independent bindings: Two parallel ways to put the separated geons back together again

JIM3s 5th layer comprises two separate parallel components. These are both concerned with combining inputs arising from the feature maps in the first three layers. So in both of these parallel components the geons which were separated in layers 1 to 3 are 'put back together again' into two different representations of the single whole object. However, these two subsystems are distinguished because they accomplishing binding of the output of the feature maps in very different ways, and the resulting representations are also very different. It is these two components of layer 5 which are the two locations in JIM3 that were chosen to change and hence implement models of less developed object recognition abilities found in adolescents and younger children (see figure 4).

3.5 The view-independent subsystem

A view independent subsystem called the Independent Geon Array (IGA) acts to form representations of explicit relations between geons - thus dynamically (but slowly!) forming a view independent structural description of the object. It accomplishes binding of the geons which result from the first three layers by identifying how individual geons relate to each other in terms of relative size and relative position within the overall object they originated from. So this attention-requiring component of layer 5 is a serial mechanism rather than the global parallel and distributed processing mechanism that operates in the view dependent surface map.

This subsystem achieves several important outcomes not achieved by the faster view dependent system. First, the attribute-relation structure is formed explicitly. Since relations among geons are made explicit they allow humans to be able to appreciate relational similarities between objects independently of whether similar object parts stand in corresponding relations. So we can appreciate two objects are similar if they have a large geon above a small geon, whatever the non-accidental properties of any of the geons. Second, relations are dynamically bound to the geons they describe. So this provides the potential for recognising complex multi-geon objects with a variety of interrelationships between the geons - to do this with static binding mechanisms such as templates might involve an impractically large set of templates ([10], page 204). Thirdly, forming relations which are invariant with geon identity and viewpoint allows the formation of a structural description that will remain the same under translation, scale and left-right reflection and is relatively insensitive to rotation in depth [9].

3.6 The view-dependent subsystem

The soonest to complete is the surface map representation in the other subsystem in layer 5. This accomplishes a view dependent static binding of geons by coding where each geon is fixed at a specific position in a Holistic Surface Map. This 2D representation captures the interrelation of geons as they were perceived in one particular view. The mapping from the output of the feature maps in the first subsystem preserves the topological relations of the geon attributes but discards their absolute sizes and location in the image. This means that the target image representation in the Holistic Surface Map is invariant with translation and scale. However, because the topological relations that are preserved in the Holistic Surface Map come from only one particular view of the object this representation is sensitive to rotation in depth and the picture plan and left-right reflection ([8], page 498). Although this second subsystem in layer five does not form structural descriptions it does have the advantage of being much faster as it does not need to wait for its inputs to include temporally separated geons, a process which takes time and can include errors.

3.7 Layers 6 to 8 Learning about multi-geon objects and recognising them when learnt

The 6th layer to 8th layers constitute the models long term memory. A simple kind of unsupervised Hebbian learning is used to encode the patterns of activation generated in layer five. Each unit in layer 6 learns to respond to geon shape attributes and relations. Units in layer seven sum input from layer six to reconstruct patterns representing geons and relations into complete structural descriptions of whole objects. These layer 7 units then activate object identity units in layer 8.

4 Simulation results for experiments using animal and artefact stimuli

4.1 Procedure for Simulation Experiments

To simulate the results from the animals and artifacts experiment of [12] we developmentally regressed JIM3 by changing two properties of the model. Figure 4 shows that the locations where the two parameters which were changed were both in layer 5 of JIM3. The parameters chosen to make less mature 'child' versions of JIM3 were the numbers of 'neurons' involved in processing in these two components. It was assumed that at earlier levels of development there might be either less resources given to recognition tasks (or perhaps these resources would be used less effectively) and this would be expected to decrease performance.

First, on the assumption that children have a less metrically-precise holistic representation of object shape than do adults, we reduced the number of locations in the model's surface map from 17 (the center plus two radii and eight orientations away from the center) to nine (the center plus two radii and four orientations); five (the center plus one radius and four orientations); and one (a single central location). Second, on the assumption that children are generally much less relational than adults in their thinking (an assumption for which there is a great deal of empirical support [5]) we removed relation units from the model's Independent Geon Array (IGA) for the child simulations. As a result of this change, the 'child' version of the model has an implicit representation of an object's inter-part relations in the surface map at an adult level, but less resources given to an explicit representation of those relations.

Before these developmentally regressed versions of JIM3 were used, we decided upon a performance measure which would allow straightforward comparison between the performance of JIM3 and the results reported by Jüttner et al.'s experiments with human participants [12]. We also developed a set of stimuli which was calibrated in a similar manner to the calibration carried out in the empirical studies with humans.

4.1.1 Performance Measure

In the original experiments of [12], human subjects (adults and children of various ages), were tested for their ability to choose the correct picture of an animal or an artifact from a display depicting an unaltered picture of that animal or artifact along with two distracters. There were two main conditions arising from use of two different types of distracter - a variant constructed by changing one part of the original object and another variant created by changing the relative part sizes of the original object (and thus effectively changing the metric relations among the object's parts).

JIM3 is not capable of performing this 'choose the correct object out of three' task (instead, it simply views one object at a time and attempts to find the best match in its LTM). Therefore, we developed a performance measure to estimate how well it would perform the choice task based on how well each object matched the correct (trained) object and each of the distracters activated the trained object's representation in the model's LTM. This measure was based on the model's response time to recognize an object (the number of iterations until an object [trained object or distracter variant] activated the corresponding trained object's representation in LTM to criterion; [8]). A second possible measure which might be used when the model could not activate the corresponding trained object representation was model's accuracy (i.e., the likelihood that an object [trained or distracter] would activate the corresponding trained object's representation in LTM). However, this was not used because the simulations typically 'recognized' both target objects and distracters as the target object, with the only distinction between conditions how many simulation cycles this took (since the distracter objects were not present in the set of recognition targets present in the learning phase).

The logic of these measures is that the more closely a distracter matches the representation of a trained object in LTM the more difficult it would be for the model to correctly reject that distracter in favour of the trained target. Accordingly, our RT-based measure of performance consists of the model's RT to 'correctly' recognize a distracter (either NAP or size change) as an instance of the trained target So although the model did not correctly reject the distracters (even very long durations eventually resulted in recognition of the learned target), it is the closest performance measure to a 'rejection' of a distracter that the current implementation of JIM3 can support.

A drawback of this performance measure is that since it compares human performance accuracy with simulation timing it does not provide a straightforward comparison between different types of task that take different amounts of time to be carried out within the simulation. This applies to the upright and inverted stimuli tasks - with inverted stimuli taking longer to be recognised than upright stimuli. This does not of course mean that inverted stimuli are easier to recognise. So within a manipulation this performance measure does allow for comparisons, but between manipulations we cannot say that longer to recognition in JIM3 infers better discrimination performance.

4.1.2 Calibration of stimuli for equal difficulty with the adult version of JIM3

The original behavioural experiments involved a calibration stage where part-change and configural change stimuli sets were formed to be of equivalent difficulty. Following this original design, we ran pilot simulations with JIM3 to equate the discriminability of the NAP and size-change variants of the trained stimuli.

Specifically, we made 5 novel multi-part part objects and trained JIM3 to recognize them, along with the dozen or so objects it was trained to recognize in the simulations reported in [8]. We then made two variants of each trained stimulus. An NAP distracter was made by changing one non-accidental property of one geon in the corresponding trained object; and a size-change distracter was made by changing the size of one geon in the corresponding trained object. During piloting we made several variants of each size-change distracter and chose, for the final simulations, the variant whose discriminability from the corresponding trained object most closely matched that of the NAP distracter. That is, following the original experiment, we explicitly equated the NAP and size-change distracters for their discriminability from the corresponding trained objects to adults. For the adult version we used JIM3 in its original 2001 version [8], with σ (the standard deviation on the Gaussian receptive fields of the memory units in layer 6) set to 0.5 and the Metric Precision and Surface Map precision set to maximum (adult) values. In figure 5 we can see in data points emphasised with dashed circles that the performance measures for the NAP changes averaged at 10.94 simulation cycles and were 10.8 for the relative size configural changed stimuli.

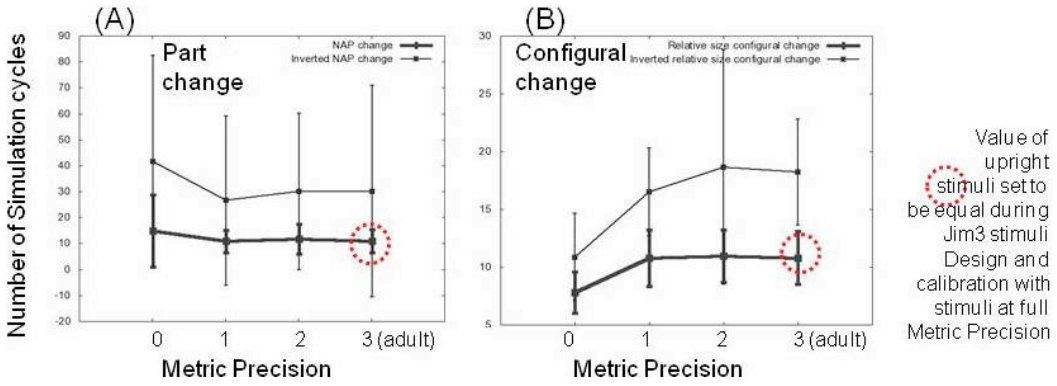

Figure 5. Showing simulation results of animals and artefacts experiment with Metric Precision at four different levels of 'development'.

4.2 Results of simulation of Animals and Artifacts Experiment

Figures 5 and 6 show the results of the two sets of computational experiments with JIM3 developmentally regressed from adult levels (3) to three lower levels of development (level '0' being the most regressed) - with figure 5 presenting results with metric relation precision in the IGA decreased and figure 6 with surface map precision decreased. Both these graphs show adult results in the upright condition circled with a dashed line - denoting that they were calibrated to be similar in value.

Figure 5 shows that as metric precision in the IGA is decreased there is a different pattern of results in the NAP change and configural change conditions. As metric precision tends to zero neurons being used, recognition performance in the configural change condition drops. However, we seem to see a performance increase with the NAP change condition when metric precision is decreased. So these simulated results show the same qualitative pattern found in the empirical results presented by [12].

Figure 6 shows that as surface map precision is decreased there is no evidence to suggest a different pattern of results between the NAP change and configural change conditions. This pattern of results is therefore different to the empirical results presented in [12]

4.2.1 Statistical analysis for MP manipulated architecture

The simulation data were analyzed with a 4 (Metric Precision level: MP level 3 (adult with 45 neurons from three receptive field classes) vs MP level 2 (30 neurons from 2 receptive field classes) vs MP level 1 (15 neurons from 1 receptive field class) vs MP level 0 (no neurons) x 2 (Manipulation: Part change versus Relative size change) x 2 (Orientation: Upright vs Inverted) mixed ANOVA with Metric Precision level as the between factor. The analysis yielded significant main effects for Manipulation [$F(1, 799) = 41.08$, $p < 0.0005$] and Orientation [$F(1, 799) = 84.56$, $p < 0.0005$] but not for Metric Precision Level [$F(1, 799) = 0.571$, $p = 0.634$.

Significant interactions were found between Metric Precision Level and Manipulation [$F(3, 799) = 5.41$, $p = 0.001$] and between Orientation and Manipulation [$F(1, 799) = 24.773$, $p < 0.005$]

Two post-hoc independent-samples t-tests were conducted to explore the interactions:

- a first independent samples t-test was conducted to compare the two most developmentally separated metric precision levels for the relative size change manipulation upright condition: MP level 3 (adult with 45 neurons from three receptive field classes) vs MP level 0 (no neurons). There was a significant difference in scores for adult MP level 3 (adult) and MP level 0; t (98) = 7.28, $p < 0.0005$, two tailed). The magnitude in the difference of the means (mean difference = 3, 95% Cl: 2.18 to 3.82) was large (eta squared = 0.353).;
- a second independent samples t-test was conducted to compare the two most developmentally separated metric precision levels for the NAP change manipulation upright condition: MP level 3 (adult with 45 neurons from three receptive field classes) vs MP level 0 (no neurons). There was a non-significant difference in scores for adult MP level 3 (adult) (M = and MP level 0; t (98) = -1.947, p = 0.054, two tailed). The magnitude in the difference of the means (mean difference = 4, 95% Cl: -8.072 to 0.77) was large (eta squared = 0.85).

4.2.2 Discussion for MP manipulated architecture

The analysis showed that there is a significant difference between relative size changed and NAP changed stimuli - but this main effect may not be a clear match to the required discrepancy between relative size changed and NAP changed conditions specified in section 1. This is because the inverted results may produce much of this main effect difference and the difference between simulations of upright stimuli may not be significant when considered on their own against each other. So a post-hoc test, discussed below, provides a finer detailed analysis of the relative size change and NAP change manipulations in the upright condition.

There is also a significant main effect of orientation between upright and inverted stimuli - with the inverted stimuli taking longer to be incorrectly recognised. Our performance measure suggests that within the same task, taking longer to be recognised is equivalent to a more accurate recognition. But between tasks this relationship does not hold. So since the simulation will actually take longer to recognise inverted stimuli because they are upside down this main effect does not show that inverted stimuli are easier to discriminate.

There was no main effect for metric precision level. However, this does not mean that there were not differences between the simulated age ranges. A significant interaction was found between met-

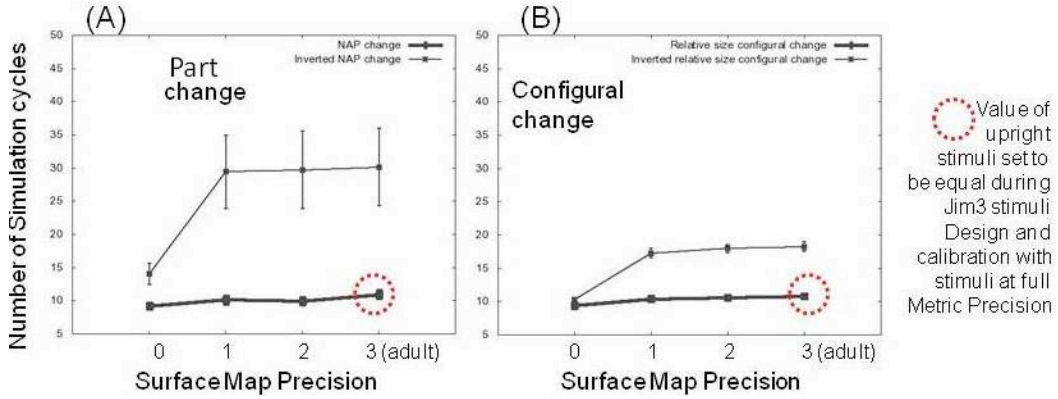

Figure 6. Showing simulation results of animals and artefacts experiment with the Surface Map at four different levels of 'development'.

ric precision level and manipulation. So as the simulation parameters modelled 'younger' parameters in the relative size change condition, performance decreased and in the NAP change condition performance increased. So these results do match the empirical results, but as noted above - the larger component of this difference between manipulation conditions may have come from the inverted results - as these mean values differ more widely than the upright conditions. The interpretation that the inverted conditions provide most of the difference between manipulation conditions is strengthened by the significant interaction between orientation and manipulation, with inverted relative size changed stimuli having the longest number of simulation cycle to recognition (in the MP = 0 condition over 40 cycles).

The complications in the analysis of considering upright and inverted orientations together was resolved with a post-hoc t-test which only looked at upright results to consider whether the 'youngest' MP regressed condition was significantly different to the adult performance level. This gave a very clear result, that whilst the relative size change condition showed lower recognition performance, than the NAP change condition was not significantly different (see figure 7).

4.2.3 Statistical analysis for SM manipulated architecture

The simulation cycle data were analysed in a 4 (Surface Map level: SM level 3 (adult with 17 neurons in two further orientations from the center neuron) vs SM level 2 (9 neurons in two further orientations from the center) vs SM level 1 (5 neurons in one further orientation from the center) vs SM level 0 (1 neuron with no further orientations from central neuron) x 2 (Manipulation: Part change versus Relative size change) x 2 (Orientation: Upright vs Inverted) mixed ANOVA with Metric Precision level as the between factor. The analysis yielded significant main effects for Manipulation [$F(1, 799) = 14.42$, $p < 0.0005$] and Orientation [$F(1, 799) = 71.81$, $p < 0.0005$] and for Surface Precision Level [$F(1, 799) = 6.4$, $p < 0.0005$].

Significant interactions were found between Manipulation and Orientation [$F(1, 799) = 16.13$, $p < 0.0005$] and between Surface Map Precision and Orientation [$F(1, 799) = 24.773$, $p < 0.001$]. The interaction between Surface Map Precision Level and Manipulation was not significant.

Two post-hoc independent-samples t-tests were conducted to explore the interaction:

- a first independent samples t-test was conducted to compare the two most developmentally separated surface map precision levels for the relative size change manipulation upright condition: SM level 3 (adult with 17 neurons) vs SM level 0 (1 neuron). There was a significant difference in scores for SM level 3 (adult) and SM level 0; t (98) = 2.81, p = 0.006, two tailed). The magnitude in the difference of the means (mean difference = 1.24, 95% Cl: 0.36 to 2.11) was moderate (eta squared = 0.074).;
- a second independent samples t-test was conducted to compare the two most developmentally separated surface map precision levels for the NAP change manipulation upright condition: SM level 3 (adult with 17 neurons) vs SM level 0 (1 neuron). There was a significant difference in scores for adult SM level 3 (adult) (M = and SM level 0; t (98) = 2.31, p = 0.023, two tailed). The magnitude in the difference of the means (mean difference = 1.74, 95% Cl: 0.24 to 3.23) was moderate (eta squared = 0.047).

4.2.4 Discussion for SM manipulated architecture

The analysis showed that there is a significant difference between relative size changed and NAP changed stimuli - but as in the MP changed architecture, the SM changed inverted results may produce much of this main effect difference. So a post-hoc test, reported below, provided a test of this point.

As with the MP regressed architecture, there is also a significant main effect of orientation between upright and inverted stimuli in the SM regressed experiments. The same explanation applies here as above - the inverted condition involves a different task so we cannot conclude inversion increases recognition performance.

There was a main effect for surface map precision level. As the simulation modelled 'younger' versions, performance levels decrease. Again, as with the MP regressed architecture, the SM-changed simulations show an interaction between surface map precision and orientation - with the longest number of simulation cycles recorded in the inverted condition. There is not a significant interaction between surface map precision and manipulation

The post-hoc t-test results highlights that the surface map regressed results do not match the empirical results reported in [12]. This test looked at whether the upright results for adult and the 'youngest' SM regressed condition were significantly different to the adult performance level. Both manipulation conditions were significantly lower performing in the youngest SM condition than the adult

condition, with a similar effect size. This pattern of results is clearly different to the empirical results reported by [12].

Figure 7 highlights the results of the post-hoc t-tests for both the MP regressed and SM regressed architectures.

Metric relation regressed	**MP 0**	**MP 3 (adult)**	**Δ**	**effect size**
Part (NAP) change	14.94	10.94	+4*	large*
Config change	7.8	10.8	-3	large
Surface map regressed	**SMP 0**	**SMP 3 (adult)**	**Δ**	**effect size**
Part (NAP) change	9.2	10.94	-1.74	moderate
Config change	9.56	10.8	-1.24	moderate

Figure 7. Key comparisons from post-hoc t-tests. This table presents results from the two computational experiments, one which simulated the developmental regression of metric relation precision (top half of table) and the other experiment which regressed surface map precision (* not a significant difference, p = 0.054)

5 Conclusions

This paper shows that recent empirical results presented by Jüttner *et al* [12] can be explained in terms of dual process models of object recognition. Simulations with the JIM3 artificial neural network suggest that a non-attentive process develops early in humans and allows part-based recognition at adult levels with the children in 7-10 age range. According to this dual process explanation, the observed developmental delay in the relative size change stimuli results from the later development of attention requiring processes that support perception of relations between object parts and the production of structural descriptions in object perception and recognition.

Removing neurons from the non-attentive surface map in JIM3 did not cause a significant difference to appear in JIM3's performance on the part (NAP) change and configural (relative size) change conditions. However, a notable and surprising result was that it took reducing the neurons all the way to zero in the attention requiring IGA to bring about a significant difference between these experimental conditions in the other set of computational experiments with JIM3. The psychological inferences that can be taken from this finding are discussed in more detail below. However, just viewing this result from the perspective of processing with machine representations provides a key lesson for articial systems engineering. This is that the dual processes in JIM3 interact together in producing behaviour so that deficiencies in attention requiring processes were masked by non-attentive processes. This highlights a more general challenge in empirical research on the structures used to represent reality - how should experimentalists untangle the interacting effects linked to multiple representation types?

The purpose of running simulations with varying precision levels for metric-relations in the IGA and the holistic surface-map was to see if either of these simulations captured the pattern of results shown in empirical observation of humans. What the human results from [12] showed was that performance for younger participants on configurally changed stimuli decreased compared with adult levels whereas performance on NAP changed stimuli stayed the same. A successful simulation should therefore show equal performance between stimuli distracter types for 'adult' parameters and show a lower performance on relative size change stimuli than part change stimuli for developmentally younger simulation parameters. As can be seen comparing figures 3 and 5, the simulations where metric-relation precision level changes in the IGA were decreased provide a good qualitative fit to the pattern observed with Jüttner et al.'s artifact and animal stimuli [12]. Since the human participants performed a different task than did the model, it is impossible to provide a precise quantitative fit between the empirical and simulation data.

The limitations in this particular modelling exercise using JIM3 are of four types. Firstly, the task that the simulation carried out was probably more simple than various strategies likely used by the human participants to eliminate distracters. In the JIM3 experiments time to recognition is always taken for stimuli presented on their own. The 'choose 1 from 3 task' gives more potential for using complex memory retrieval strategies than simply measuring time to recognition for a single object. In addition, which strategies might be used in either task is likely to change through development independently of the changes to resources given over to metric relation or surface map precision. Developing proficiency in metacognition and increasing cognitive resources have been presented as competing explanations in memory development [6]. So we might expect analogous competing theories of 'increasing metacognition' and 'increasing cognitive resources' when attempting to explain developmental trajectories in object recognition.

Secondly, the images that JIM3 learns and then recognises are simpler than the naturalistic 2D images used by [12]. The naturalistic images possess difference in texture and colour which the stimuli used by JIM3 do not possess.

Thirdly the modelling exploration has been set up as a two horse race, to decide which of these changes to JIM3 provides the best fit for the pattern of empirical results for adults and children described by [12]. Each of these regressions was 'clean' in the sense that only one parameter at a time was regressed. In a real infant we might expect both MP and SM precision to decrease as well as there being a number of other changes that involve lower recognition performance for younger participants. For example, on the assumption that children have less stable and/or precise memories for objects than do adults, we might change σ on the Gaussian receptive fields in layer 6 of JIM3 from 0.5 (the value in the adult simulations) to 1.0. This increase would have the effect of making any given unit in Layer 6 more tolerant of deviations from its preferred pattern (corresponding to the center of the distribution). Possible future computational experiments with JIM3 might therefore involve co-varying the two existing changes with each other and with changes in σ. However, preliminary experiments have shown that decreasing σ on its own does not cause relative size stimuli to be processed less effectively - with mean simulation runs actually higher for relative size stimuli at a value of σ that gives minimal recognition performance.

Lastly, both the empirical results and the modelling research do not rule out the impact of differing life experience and consequent encoding differences in memories might have in the performance of JIM3 after layer 5.

These four limitations of: (1) task and strategy simplicity; (2) 'clean' changes to parameters; (3) image simplicity; and (4) learning experience in the simulation being equal between regressed and adult architectures; might all be expected to increase recognition performance in JIM3 compared with human performance. So it may be as a result of a combination of these factors that it took decreasing the metric precision neurons to zero to get a large drop in performance. Alternatively, the finding that only the 'MP=0' condition provides a large decrement in performance may be suggesting that children of

age 7 to 9 years old really do have a much lower than previously expected ability to make metric judgements in visual object recognition. That this is not apparent in day to day life or in other kinds of object recognition experiment may be because this lower ability will only be apparent when children view objects in a way that their highly performing 2D systems cannot quickly produce recognition. Otherwise partial-orderings rather than absolute metric judgements may suffice. So one suggestion for future work is to adapt JIM3 so that it can support more complex tasks and more complex strategies, with image simplicity matched, with many parameters being systematically changed during simulations, and with learning regimes matched to those that the adults participants experienced. Some of these suggestions have already been carried out - for example experiments have been conducted which control for differing previous experience with novel objects - see [12] experiment 3. The finding that JIM3 needs to have no metric relation precision to qualitatively match 7-9 year old human performance might also suggest new empirical studies where participants learn novel objects but are then presented with very different views of these objects so that the view dependent system would not be expected to maintain high performance levels.

There are also a number of deeper issues linked to the core features of JIM3. For example, in JIM3, both the view dependent and view independent routes through the architecture use geons as a fundamental representational unit. However, it is not a settled issue what the basic level in structural descriptions in visual object recognition are. For example, children from 3 to 4 made less use than adults of the shape boundaries that distinguish different types of geons [1]. So to model children's performance we might want to relax the requirement that geons are a fundamental representational unit at earlier stages in development. In addition, it is also worth noting that JIM3 possesses surfaces in layer 3 of the architecture, but these surfaces are only used in the assignment of geons before layer 4 - rather than primitives for the spatial relationships recorded in the view independent component of layer 5. However, surfaces have been proposed as representational primitives within spatial relationships [13] .

In addition, in JIM3 there is limited opportunity for processing in later layers to influence earlier processing in an on-line dynamic fashion. For example, top-down effects of memory on processing before layer 5 through backward projections do not occur in JIM3. We might imagine that attention emerges over moment to moment as an internal representation of an object emerges - a dynamic process not captured within JIM3. Instead, in JIM3, attention is 'on full' as the object starts to be represented.

Lastly, JIM3 is a dual process model where each process is supported by different hardware, in the form of separate neural networks in layer 5. Other dual process theories have a similar arrangement. For example, object perception and action are proposed to occur in two separate dorsal and ventral streams [7]. Alternatively, the idea of dual processes can be de-linked from the idea of dual 'systems'. It may be different processing occurs at different times on a common substrate. So 'dual process - one system' could be a design schema for a new object recognition system where compositional and non-compositional processes are separated in time but not space.

So in summary - a version of JIM3 with regressed metric relation precision in the IGA has been shown to provide a better match to empirical results than a regressed holistic surface map version. An interesting finding is that even small numbers of neurons present in the IGA can provide similar level of recognition performance to an 'adult' JIM3 with its full complement of neurons. Though the lessons for human psychology from this are still to be worked out this work does provide an example for research in machine representation of the benefits of dual representation systems. Future work has also been suggested that: (1) would involve adapting JIM3 to more closely match the types of task and stimuli and learning pattern used in empirical studies of object recognition development; (2) that would involve empirical testing of younger adolescents with stimuli that have been rotated so that the view dependent mechanisms do not provide an effective route to recognition; and (3) would involve developing alternatives to JIM3 that support surfaces as a representational primitive, provide more backward projections to provide top down effects of existing knowledge, and development of dual process-single system models where differences in processing exist across time but not across resources.

Philosophers have long theorised about compositionality and its benefits. This research illustrates the challenges in investigating how object representations develop. These include that, in natural systems, there is no transparent access to internal representations, multiple representational forms can interact to produce complex behavioural patterns, and the existing implemented computational models do not always neatly fit completely with emerging empirical paradigms. However, computational modelling can nonetheless show how neural systems can support representational diversity in humans, other animals and machines.

REFERENCES

[1] M. Abecassis, M.D. Sera, A. Yonas, and J. Schwade, 'What's in a shape? children represent shape variability differently than adults when naming objects', *Journal of Experimental Child Psychology*, **78**, 213–239, (2001).

[2] I. Biederman, 'Recognition by components: A theory of human image understanding', *Psychological Review*, **94**, 115–147, (1987).

[3] J. Davidoff and D. Roberson, 'A theory of the discovery and predication of relational concepts', *Journal of Experimental Child Psychology*, **85**, 217–234, (2002).

[4] L. Doumas, K. Holyoak, and J. Hummel, 'The problems of using associations to carry binding information', *Behavioral and Brain Sciences*, **29**, 74–75, (2006).

[5] L. Doumas, J. Hummel, and C. Sandhofer, 'A theory of the discovery and predication of relational concepts', *Psychological Review*, **115**, 1–43, (2008).

[6] J.H. Flavell, 'First discussant's comment: What is memory development the development of?', *Human Development*, **14**, 272–278, (1971).

[7] M.A. Goodale and A.D. Milner, 'Separate visual pathways for perception and action', *Trends in Neurosciences*, **15**(1), 20–25, (1992).

[8] J. Hummel, 'Complementary solutions to the binding problem in vision: Implications for shape perception and object recognition', *Visual Cognition*, **8**, 489–517, (2001).

[9] J. Hummel and I. Biederman, 'Dynamic binding in a neural network for shape recognition', *Psychological Review*, **99**, 480–517, (1992).

[10] J. Hummel and B. J. B.J. Stankiewicz, 'Categorical relations in shape perception', *Spatial Vision*, **10**, 201–236, (1996).

[11] M. Juttner, A. Muller, and I. Rentschler, 'A developmental dissociation of view-dependent and view-invariant object recognition in adolescence', *Behavioural Brain Research*, **175**, 420–424, (2006).

[12] M. Juttner, E. Wakui, D. Petters, S. Kaur, and J. Davidoff, 'Developmental trajectories for part-based and configural object recognition in adolescence', *Developmental Psychology*, **48**, Online first publication March 26, 2012, No pagination specified, (2013).

[13] E.C. Leek, I. Reppa, and M. Arguin, 'The structure of three-dimensional object representations in human vision: evidence from whole-part matching.', *Journal of experimental psychology. Human perception and performance*, **31**, 668–84, (2005).

[14] I. Rentschler, M. Juttner, E. Osman, A. Mller, and T. Caelli, 'Development of configural 3d object recognition', *Behavioural Brain Research*, **149**, 107–111, (2004).

[15] F. van der Velde and M. de Kamps, 'Neural blackboard architectures of combinatorial structures in cognition', *Behavioral and Brain Sciences*, **29**, 37–108, (2006).