Using Computational Models of Object Recognition to Investigate Representational Change Through Development

Dean Petters, John Hummel, Martin Jüttner, Elley Wakui and Jules Davidoff

Abstract Empirical research on mental representation is challenging because internal representations are not available to direct observation. This chapter will show how empirical results from developmental studies, and insights from computational modelling of those results, can be combined with existing research on adults. So together all these research perspectives can provide convergent evidence for how visual representations mediate object recognition. 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

D. Petters (✉)
Birmingham City University, Birmingham, UK
e-mail: dean.petters@bcu.ac.uk

J. Hummel
University of Illinios at Urbana-Champaign, Champaign, USA

M. Jüttner
University of Aston, Birmingham, UK

E. Wakui
University of East London, London, UK

J. Davidoff
Goldsmiths College, London, UK

G. Dodig-Crnkovic and R. Giovagnoli (eds.), *Representation and Reality in Humans, Other Living Organisms and Intelligent Machines*, Studies in Applied Philosophy, Epistemology and Rational Ethics 28, DOI 10.1007/978-3-319-43784-2_8

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

Only a fraction of a second passes from when a person sees a familiar visual object to when they can then name it. Despite it being relatively quick, the process of visual object recognition is complex, with multiple sub-processes, some occurring in parallel. Multiple forms of representation are invoked in object recognition, from the point of initially perceiving an object to finally being able to provide the name for it. A very rough framework to start understanding the complexities of recognition is to consider how recognition processes get started and how they complete. First, a person perceives an object is present in sense data. Then this perceptual pattern is compared with a representation of some kind for that object type in long-term memory. When a match is found between these two representations of the object currently present in perception, recognition has occurred. However, this brief sketch is an oversimplification because perceptual processes do not complete before memory retrieval starts. Rather, perception and memory retrieval occur together and influence each other. In addition this rough distinction between perceptual and memory processes leaves open many further questions. For example, can the perceptual component of the overall recognition process be broken down to more basic sub-processes?; how are visual features in sense data selected and bound [44, 48]?; how are part-relations bound [28]?; and, what is the role of attention in object recognition [34]? Questions also arise from considering how percepts are matched to memories: does object recognition rely on one, two or many mediating representations, in perception, and in long-term memory [21]?; and what backwards projections or 'top-down' influences from memory retrieval to perception exist [19]?

The main question that this chapter is concerned with arises from the finding that object recognition performance changes through adolescence. So this chapter explores possible developmental trajectories for how object recognition representations change during this period in development. Specifically, this chapter will use computational modelling to attempt to explain empirical evidence for differences in the visual representations used for object recognition in middle childhood (ages 7–10) and adulthood. In attempting to answer this specific question, some general issues in recognition processes and representations for recognition will need to be examined. So before the developmentally focussed question can be answered we will consider what representations are used when adults recognise objects. Visual object recognition has been far more intensively studied in adults than children, and current theories propose that adults use a variety of representations when recognising objects. These include compositional representations which describe objects as three-dimensional (3D) structures in terms of the interrelationship of their parts

[4, 29, 36], and image-based representations which capture two-dimensional (2D) object views [18, 37].

1.1 Compositionality in Visual Object Representations

Humans are highly accomplished at recognising objects visually. Familiar objects can be recognised in novel viewpoints, and new unseen members of familiar categories are also often recognised with speed and accuracy. Structural description theories explain this impressive performance by proposing that recognition of objects occurs through intermediate object representations that are compositional in nature and are abstracted from sensory data. Formal logics and natural languages demonstrate compositionality because the meaning of linguistic or logical expressions with multiple parts is determined not just by the meaning of those parts but the way they are put together. In addition to language, compositionality is found in a diverse range of other entities in the world including visual objects. So in a recognition task an object's identity can be determined by identifying relations between its component parts, not just the nature of those parts viewed in isolation [4]. In our interactions with objects the perception of compositionality can be manifested across multiple modalities [32]. We can perceive visual compositionality in scenes and objects and thus form structural descriptions. Structural description recognition processes that are mediated by compositional representations are also termed analytic processes because they specify the relations among an object's parts explicitly and independently ([43], p. 257).

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 for objects [4]. Geons can therefore be compared with 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 [4]. 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. For example, consider a typical coffee mug. The spatial relation between the handle and the body of a coffee mug might be explicitly described as a small curved cylinder side attached to a vertical straight cylinder. This structural description would match a whole range of slightly different coffee mugs. So what structural descriptions allow is a generalisation across metric variations that are still within categorical divisions. This property of structural descriptions is related to the issue of view-invariance to rotation discussed in the next section.

Artificial neural networks can represent visual compositionality and hence model natural cognition [21, 47]. 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 [12, 47], and how generalised shape information develops [13].

1.2 Non-compositional Image-Based Object Representations

Visual object recognition that is accomplished through the use of image-based representations provides a complementary capability to recognition relying on compositional representations [21]. This is because abstracting images into parts and relations between parts takes time and is potentially error prone. Avoiding these drawbacks, and just using the actual image in sensory data to match against a similar type of unabstracted view-based representation for objects in long-term memory can therefore afford a faster and potentially less error-prone route to recognition. Image-based recognition will also be advantageous when an object does not possess distinct separable components which can be described compositionally, such as is the case with a cloud or piece of wire. Image based representational theories propose that the particular 2D object views received in sensory data are encoded in memory. This is an opposing account to structural description theories because no abstract representations are held in memory. View-based representations are also described as 'holistic' because, although they may include visual features or fragments, these are not represented independently but rather at fixed locations within the 2D object overall shape. In terms of compositionality, such image-based representations do not get decomposed into recombinable components. So view-based (holistic) coding is analogue rather than compositional because object parts that are big in the image have to be represented as a big proportion of the view-based representation, if one part is above another in the image it has to be positioned above in the representation, not just described as such. In view-based representations object parts are represented in terms of their topological positions in a 2D space described by the outline of the object [39, 41]. When perceptual images are matched to view-based representations in memory this occurs "all of a piece" [43]. This means that an object is recognised according to its overall shape and features within the image but not in terms of the interrelationship of isolatable parts.

To recognise an object such as a horse, a view-based representation of a horse in memory would be matched "*in its entirety*" against the object's perceptual image to determine the degree of fit [39, 41]. As Thoma, Hummel and Davidoff note, the "*process is directly analogous to laying a template for* [a] *horse over the image of the* [] *horse and counting the points of overlap*" ([43], p. 259).

1.3 Differing Predictions for Performance Outcomes from Rotation in Depth

When a known object is presented from an unfamiliar view the 2D view-based representation in memory may not fully match up to the currently sensed image because rotated objects can present different 2D images from the familiar views that are already in memory. Image-based theories propose that known objects seen in unfamiliar views can still be matched and recognised by bringing these perceptual images into line with stored images in memory. This is postulated to occur through processes of normalisation, which can include alignment, mental rotation and view interpolation [8, 30, 31, 37, 38, 45, 46]. An important characteristic of all these normalisation theories is that they predict a linear cost in recognition performance due to rotation. Structural description theories do not make the same prediction of linear cost to rotation. Figure 1 helps illustrates the difference between predictions on rotation costs to recognition performance for view-based and structural description mechanisms. A recognition task based on this object would involve an experiment with a learning phase and a test phase. In the learning phase, a participant would

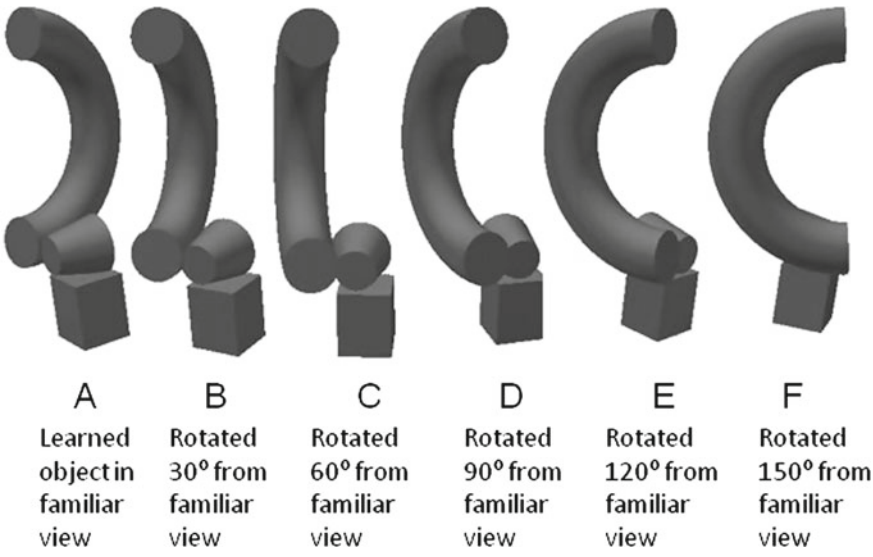

Fig. 1 Figure showing why structural description and image-based theories make different predictions about performance in response to rotation in depth. After the object has been learned in the view shown by stimuli *A*, this and other novel views (*B*–*F*) can be tested for how quickly they are recognised as the same as the familiar learned object view. View-based theories predict that recognition latencies will increase proportionately with rotation magnitude, from stimuli *B* through to stimuli *F*. Structural description theories predict no added recognition cost from stimuli *B* through to stimuli *D*, as all these stimuli are described with the same structural description, and it is the abstract structural description that is matched to memory

become familiar with object A, perhaps by learning a name or label for this object, but would never see the object in the rotated views B–F. Then in the test phase of the experiment, the participant would have to show recognition of objects, perhaps by assigning them as already learned but rotated objects or novel objects that have never been seen in any view. In the images presented in Fig. 1, a view-based process with a normalisation mechanism with linear cost to rotation would predict stimuli B would take longer to recognise than A, and C longer than B, D longer than C, through to F, but with the same added cost to recognition latencies as the magnitude of rotation increased from presentations of B to a presentation of the F view of the stimuli. However, Fig. 1 shows why changes in viewpoint will often make no difference to the structural description that results when perceiving the object from different rotated views. Despite the fact that the rotations in depth for objects B–F in Fig. 1 result in significant changes to the observable 2D image, object views for stimuli B, C and D generate exactly the same structural description as the original learned familiar object A, with the same relations between the same component parts [large curved cylinder side-attached small truncated cone] and [small truncated cone above small cube]. So for objects B–D recognising the same structural description as A would predict the same recognition latencies. Object E has the truncated cone mostly occluded, and in object F it is completely occluded. So a view-invariant response to rotation would be unlikely with these object views. In summary, structural description models predict that rotation will not affect recognition performance when the same structural descriptions will reliably result, and image-based theories predict that, when novel rotated views are perceived, recognition is not immediate and some process is required to transform the perceived 2D image to check matches against previously viewed 2D exemplars in memory.

1.4 The View Dependence/Invariance Debate

Recent object recognition theories often invoke some form of structured representation activity in parallel with image-based representations (though details differ between particular approaches) [2, 3, 17, 21, 27]. This relatively high level of agreement within the object recognition community was not always the case. In the late 1980s and most of the 1990s, there occurred a vigorous debate about the nature of internal representations for object recognition amongst psychologists carrying out visual object recognition research. On the one hand were proponents of structural description models, such as the RBC theory of object recognition [5–7]. On the other hand were those who promoted view-based theories which proposed representations much closer to the sensory input [18, 37, 38]. The differing predictions illustrated by Fig. 1 led to an approximately 10 years debate in object recognition research. The view-dependence/invariance debate assumed that object recognition performance that was invariant to object rotation provided evidence for structural description representations, and rotational costs in performance were viewed as evidence for image-based representations in recognition. The debate occurred in these

terms because researchers who were actually interested in elucidating the nature of internal representations used object recognition performance for known objects in novel rotated viewpoints as a proxy or surrogate for the nature of the internal representations believed to be involved in object recognition.

As the number of published studies increased, significant evidence was found that supported both positions in the view-dependence debate on visual object recognition. Sometimes recognition occurs with little performance cost from stimulus rotation, supporting the view-invariant predictions of the RBC theory. Other studies report a pronounced cost to rotation, with increasing performance costs as the magnitude of the rotation increased. As Hayward noted in 2003:

> the viewpoint debate appears to have run out of steam. It has ended because, on most major issues, the two sides are in basic agreement. Both agree that a change in viewpoint will normally result in viewpoint costs, albeit small in some cases. Both agree that some visual properties, particularly those related to the structure of an object, will be particularly important for generalizing across viewpoint. Finally, neither can deny the findings of the other, both view-invariant and view-specific patterns of data have been replicated so often that it has been difficult for either to argue that their opponents' results are a special case ([17], p. 425).

So a consensus has arisen that both image-based representations and structured representations with some degree of abstraction from the sensory image are involved in visual object recognition. With dual processes mediated by contrasting representations operating in parallel we can produce rotation performance predictions that have a different, more complex pattern. We can expect the fastest recognition to occur when an image is perceived in a familiar view, as both structural description and view-based mechanisms will operate optimally. Next fastest will be images that are rotated but which are still parsed to give the same structural description as the familiar view. After this, rotations that result in occlusions of geons that are clearly seen in the familiar view, or when previously unseen geons become viewable, result in the slowest recognition. The next section will present further convergent evidence for the operation of dual recognition processes. These dual processes (with dual representations) act as complementary solutions to object recognition because the dual processes draw upon the contrasting strengths of image-based and compositional representations.

2 Accumulation of Further Evidence for Multiple Representations Mediating Adult Object Recognition

2.1 Fractionating the Visual Object Recognition System into Independent Components

At the close of the view-dependence/invariance debate, the mixed results from studies of object rotation suggested that both view-invariant and view-dependent

mechanisms are likely to be operating. Stankiewicz [33], and Foster and Gilson [15] responded to this state of affairs with research aiming to observe independent dimensions of the object recognition system using experimental manipulations. These experiments thus enable us to see how structural and view information is combined by object recognition processes [17]. Foster and Gilson used 'paper clip' stimuli that did not possess geons, and that varied in their number of parts (structural information) and in view-specific properties such as length, degree of curvature and angle of joints between parts. They then assessed how each of these properties affected discrimination performance. Experiments showed object structure information and image-based information are both independent and are combined additively. Stankiewicz conducted experiments which showed that 3D properties, such as primary axis curvature and aspect ratio, are estimated independently of 2D object image properties. He also showed that 3D shape is estimated independently of object viewpoint. As Stankiewicz notes [33], the results of fractionating the visual object recognition system into independent components provide strong evidence for a dual process model of object recognition, with a view-invariant component that forms structural descriptions and hence compositional representations, and a view-dependent component that forms representations much closer to the unabstracted sensory image.

2.2 Neuropsychology Evidence for Dual Representations in Object Recognition

Evidence for dual processes (and hence dual representations) occurring in visual object recognition is also provided by neuropsychological case studies. Dual process theories of visual object recognition postulate that each process acts upon a different kind of representation. Analytic processes involve perception of object, parts before the formation of compositional representations. Holistic processes do not involve such a decomposition. Davidoff and Warrington found patients who could not recognise individual object parts but who could name whole objects [10, 11]. However, the fact that the intact naming ability only occurred in familiar object views suggests that these patients' abilities resulted from holistic recognition whilst analytic recognition no longer functioned.

2.3 Priming Evidence for Dual Representations in Object Recognition

Perhaps the most comprehensive and persuasive evidence for the dual nature of representations for visual object recognition comes from behavioural studies which measure how priming can improve recognition performance. Priming involves the

repeated viewing of an object, and priming effects help infer the nature of representations because of various changes that can be made in the manner in which the priming object can be depicted [40]. Priming benefits to recognition performance are measured as the difference in latencies between repeated and unrepeated (and so unprimed) probe images [39, 41]. The key finding from priming studies is that performance can be improved by two kinds of priming: (1) by primes that are presented so that they are attended to and (2) with primes that are presented so that they are not attended to [43]. Attended primes are believed to activate both structural description and view-based representations. Unattended primes are presented outside the focus of attention for a brief enough exposure that attention cannot switch to them. Priming images in these cases only increase recognition when they exactly match the object being recognised. However, attended primes can be altered with various modifications that leave the structurally described elements intact and still provide performance improvements. These modifications include being mirror-reflected [23], split and recombined [43], rotated in the picture plane [42] and rotated in depth [41].

Splitting images and then recombining them is an interesting modification for a prime because split and recombined images give highly similar structural descriptions if the image is not split where it would naturally be parsed into its geon components. Thoma, Davidoff and Hummel [43] use images where splitting in the middle resulted in all part shapes being recoverable in the recombined image. So split images disrupt view-specific matching, and hence result in no performance increase when unattended. However, they do give a performance increase when attended, as analytic recognition is based on discrete, isolatable parts which are still present in the split and recombined stimuli [40].

Section 1.4 explained that recognition after rotating objects in depth did not provide a clear-cut distinction between view-invariant and view-dependent processes. This is because both structural description theories and view-based theories provide explanations for how humans recognise rotated stimuli. Structural description theories posit representations that do a lot of the 'heavy lifting' in recognition tasks. So such relatively rich representations only need to be acted upon with relatively simple processes to produce flexible recognition behaviour. If image-based representations are acted upon with similarly simple processes, they cannot be expected to perform as flexibly as systems using structural descriptions. But the key issue is that image-based representations may be operated upon with complex and sophisticated processes which compensate for the lack of flexibility in the representation. So an object recognition researcher may be left using view-dependence/invariance to distinguish two kinds of representation that predict similar outcomes because structural description theories postulate simple operations on "smart" representations, whereas view-based theories postulate "smart" operations on simple representations ([20], p. 160). However, using priming images rotated in depth works better at distinguishing underlying representations because it works by priming representations that mediate view-invariant mechanisms and these representations are not involved in view-dependent mechanisms. So as with priming probes formed from split images, prime probes that have been rotated in depth still possess the same parts and part-relations as the unrotated images, and hence give a performance benefit to mechanisms that are

mediated by structural description representations. Rotating objects in depth results in significant changes to the observable 2D image, and this image-based representation has no carry-over effect on the recognition task and so does not give a priming benefit to performance.

Priming studies not only show that the structure of part-relations in priming images is only captured in attended conditions, but also that performance improvements from priming from same-view pairs is equivalent in attended and non-attended conditions (around 50 ms improvement in recognition performance when a same-view image is previously presented attended or unattended). That the priming benefit of a view-based representation is equal with and without attention suggests that the process that provides this benefit occurs independently and that the recognition system has at least two independent components [39, 41]. In summary, the results of many priming studies show that attended prime images reliably primed exactly the same view as well as many modified images that kept the prime images, structural description elements intact, whereas unattended images only primed themselves in exactly the same view [39].

2.4 Brain Imaging Evidence for Dual Representations in Object Recognition

Whilst behavioural studies using priming techniques have made great progress in consolidating knowledge of the representational distinctions in object recognition processes [39]. Thoma and Henson [40] have also extended the behavioural findings from priming experiments by conducting the first brain imaging studies to provide neural evidence for dual processes mediated by contrasting compositional and view-based representations. Their results implicate a ventral stream in attention requiring processing mediated by compositional representations and a dorsal stream implicated in view-based recognition which does not require attention. They adapted priming paradigms which used split and recombined priming stimuli. As Thoma and Henson [40] note:

> The current findings support hybrid models of visual object recognition that include both analytic and holistic object pathways, with the analytic pathway dependent on visual attention. Regions in the left ventral visual stream only showed repetition suppression (RS) from primes in more anterior fusiform regions, and the amount of this RS correlated with the amount of behavioural priming, consistent with an analytic pathway. Regions in the dorsal stream on the other hand, specifically the intraparietal sulcus, showed repetition enhancement (RE) only for intact primes, regardless of attention and the amount of RE correlated with the amount of behavioural priming from uncued, intact primes, consistent with a holistic pathway. ([40], p. 524)

2.5 Summary of Representational Properties of Dual Recognition Systems

Dual process theories have become quite well established through convergent evidence from behavioural, neuropsychological and imaging studies. In these theories, it is hypothesised that the representations mediating recognition differ as a function of whether an object is attended or ignored [35] and multiple visual representations are activated in response to attended objects. These include: (1) compositional structural descriptions with explicit relations that require attention as visual features must be bound into parts and parts then need to be bound to relations; and (2) non-compositional image-based representations of specific views that are activated in response to attended and non-attended objects (that are outside the focus of attention). In addition to being distinguished by whether they require attention, representations also differ in how much they are abstracted from the sensory image. Compositional representations need to abstract away from many details of the image present in sensory data, whilst image-based representations do not need to significantly abstract from sense data. Figure 2 presents some distinctions between compositional and view-based recognition processes.

However, the great majority of this convergent evidence for the operation of dual processes (and dual representations) in visual object recognition has come from studies using adults as participants, leaving open the possibility of different developmental trajectories for image-based and compositional representations. The next section presents relevant empirical evidence from recent studies with adults and children.

Fig. 2 Table summarising the contrasting properties of the representations mediating structural description and view-based recognition processes

Structural description representations	**View-based representations**
Abstracted away from particular examples in sense data	Based upon view - specific details of an exemplar image with no abstraction
Compositionally formulated representations in perception and memory	Image-based/analogue representations in perception and memory
Analytic (decomposition into parts and recombination with relations between parts)	Holistic (no decomposition into isolatable parts and no explicit relations between parts)
3D representations	2D representations
Response to object rotation is viewpoint invariance performance	Response to rotation is viewpoint dependence performance
Ventral stream implicated	Dorsal stream implicated

3 The Development of Configural Processing in Object Recognition: Recent Empirical Results

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 [9, 24, 32]. Recently, Jüttner et al. [26] examined developmental trends associated with identification of correct pictures when presented alongside incorrect distracters (in a 3 alternative-forced-choice (AFC) task). Two distracter types were compared: part-changed stimuli, where one part of the stimuli was substituted for an incorrect part (Fig. 3); and a change to the overall proportions of the object (the configural change condition, Fig. 4).

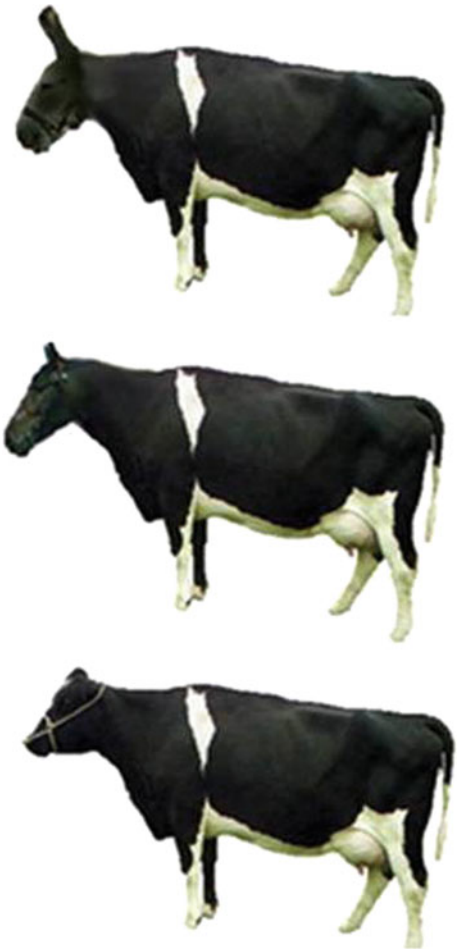

Fig. 3 Showing an animal version of a part-change stimuli used in human studies. Selecting the 'real' cow image is a non-configural task as only one object-part needs to be checked at a time

Fig. 4 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

In both part-change and configural (relative size) change conditions, the task is to choose the 'correct' image. So in Fig. 3 the bottom 'cow' is the only image with a cow's head. In Fig. 4 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 [26], p. 163 for examples). Responses to defined-base, rigid artefact stimuli (Figs. 5 and 6) showed the same pattern of results as 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

Fig. 5 Showing an artefact version of a part-change stimuli used in human studies. Selecting the 'real' bicycle image is a non-configural task as only one object-part needs to be checked at a time

and 16 years in upright and inverted conditions. Overall, 32 participants were used in each of six 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 [26]. 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-years-old) are that, whilst part-change performance is marginally lower than adult levels, relative size change performance is significantly lower. For older children (11–16-years-old), part-change performance has reached the adult level whilst relative size change performance is still not fully consolidated [26]. Figure 7 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. Developmental studies have used the differential performance of recognition using configural processing and part-based processing as a surrogate for differing access/use of image-based and structural description representations. This is because, whilst part-based recognition

Fig. 6 Showing an artefact version of a relative size change stimuli used in human studies. Selecting the 'real' motorbike image is a configural task as recognition results from checking the relative sizes of two (or more) parts

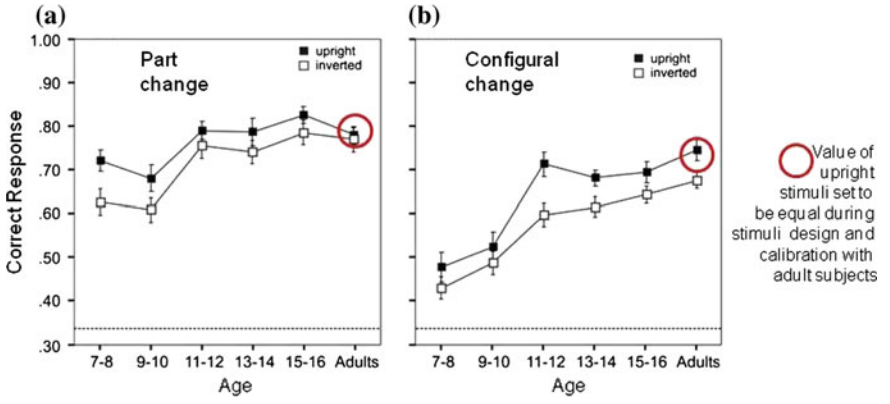

Fig. 7 Results of experiments where participants of different ages were tested with part and configural changed stimuli

only requires focussing on a single object part, configural processing requires attending to at least two parts at the same time.

To evaluate further this possible dual process explanation for these results, this paper now presents simulation results gained by developmentally regressing JIM3 [21], a prominent dual process model that simulates visual object recognition. JIM3 was created as an implementation of RBC theory as a computer simulation, and its use in computational modelling of human vision demonstrates how machine vision algorithms can investigate representations for object recognition.

4 JIM3: A Dual Process Model of Object Recognition

4.1 Introduction to JIM3

JIM3 is an eight-layer artificial neural network model of visual object recognition [21–23]. 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 8 (adapted from [21]) shows JIM3's eight layers and the two places where changes were made to developmentally regress the architecture.

4.2 Layers 1–3: From Feature Maps to Independent Geons

The first three layers comprise 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 simulates 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 [22] as fast enabling links (FELs).

The first three layers of JIM3 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

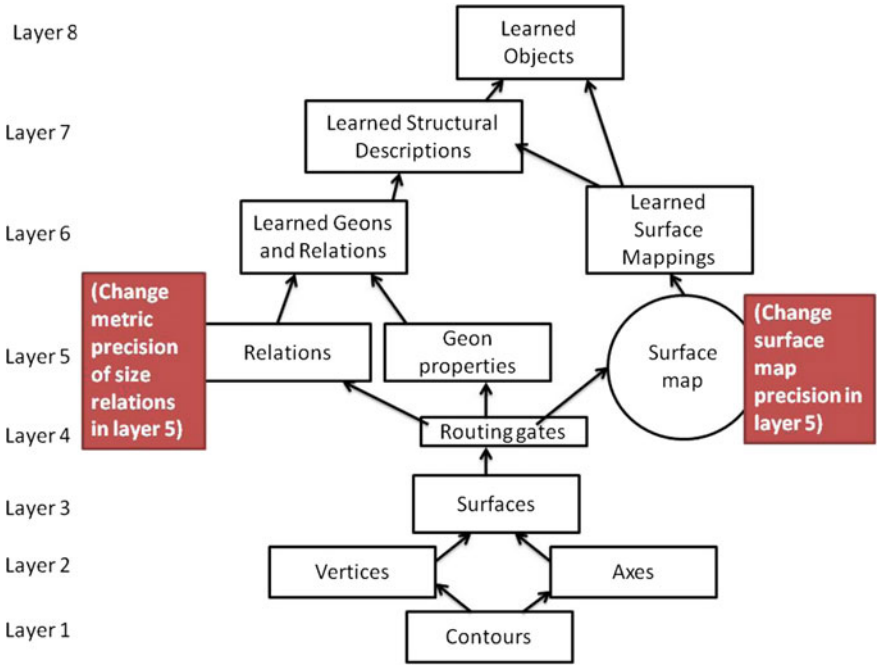

Fig. 8 Diagram of JIM3 showing the two locations in the architecture where changes were made to capture this architecture's performance for an earlier developmental stage

which cause FELs to treat units as from the same geon are: local coarse 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.

4.3 Layer 4: Routing Gates (Passing Each Independent Geon Forward Separated in Time from the Other Geons)

The fourth 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, 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.

4.4 Layer 5: View-Dependent and View-Independent Bindings: Two Parallel Ways to Put the Separated Geons Back Together Again

JIM3's fifth 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 1–3 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 Fig. 8).

4.5 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 enable humans 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 ([23], p. 204). Figure 9 presents three different objects all described by the same set of attributes:

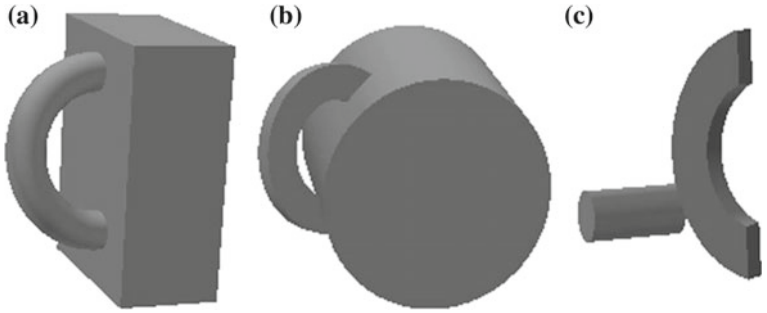

Fig. 9 Figure showing how very different objects can be formed from combination of the same attributes but with different relations between the attributes. Object **a** has a large geon with straight sides and straight cross-section beside a smaller geon which is curved along axis with a round cross-section. Object **b** has a large geon with straight sides and a round cross-section beside a smaller geon which is curved along its axis and with a straight cross-section. Object **c** has a large geon which is curved along its axis with a straight cross section beside a smaller geon with straight sides and round cross section

a large part and a small part, one part curved and one straight edged and one part with a straight cross-section and one part with a curved cross-section. What distinguishes these two objects is how particular attributes are dynamically linked to each other. So in the attributes for object (a), the large geon attribute is linked to straight sides and straight cross-section, whilst the small geon attribute is related to curved edge and curved cross-section. 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 [22].

4.6 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 and metric properties 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 and metric properties 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 ([21], p. 498). Although this second subsystem in layer 5 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.

4.7 Layers 6–8: Learning About Multi-geon Objects and Recognising Them When Learnt

The sixth to eighth layers constitute the model's long-term memory. A simple kind of unsupervised Hebbian learning is used to encode the patterns of activation generated in layer 5. Each unit in layer 6 learns to respond to geon shape attributes and relations. Units in layer 7 sum input from layer 6 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.

5 Simulation Results for Experiments Using Animal and Artefact Stimuli

5.1 Procedure for Simulation Experiments

To simulate the results from the animals and artefacts experiments of [26] we developmentally regressed JIM3 by changing two properties of the model. Figure 8 shows that the locations where the two parameters 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 centre plus two radii and eight orientations away from the centre) to 9 (the centre plus two radii and four orientations), 5 (the centre plus one radius and four orientations); and 1 (a single central location). So with 17 neurons the model's surface map provides the most precise representation of the target object metric properties and when reduced respectively to nine neurons the model loses the eight neurons that provide the most fine-grained metric precision. Then with each further reduction in surface map neurons, it is again the neurons that provide the most metric precision that are removed. 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 [13]), 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 [26]. 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.

5.1.1 Performance Measure

In the original experiments of [26], human subjects (adults and children of various ages) were tested for their ability to choose the correct picture of an animal or an artefact from a display depicting an un-altered picture of that animal or artefact 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 long term memory (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 [21]). A second possible measure which might be used when the model could not activate the corresponding trained object representation was the 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 'recognised' both target objects and distracters as the target object, with the only distinction between conditions being 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' recognise a distracter (either non-accidental-property (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.

5.1.2 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 five novel multi-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 [21]. 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 [21], 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 Fig. 10 we can see 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.

5.2 Results of Simulation of Animals and Artefacts Experiment

Figures 10 and 11 show the results of the two sets of computational experiments with JIM3 developmentally regressed from adult level (3) to three lower levels of development (level ‘0’ being the most regressed), with Fig. 10 presenting results

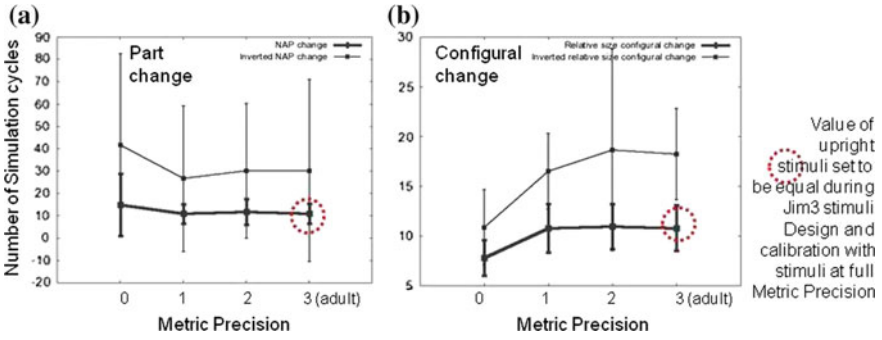

Fig. 10 Showing simulation results of animals and artefacts experiment with Metric Precision at four different levels of 'development'

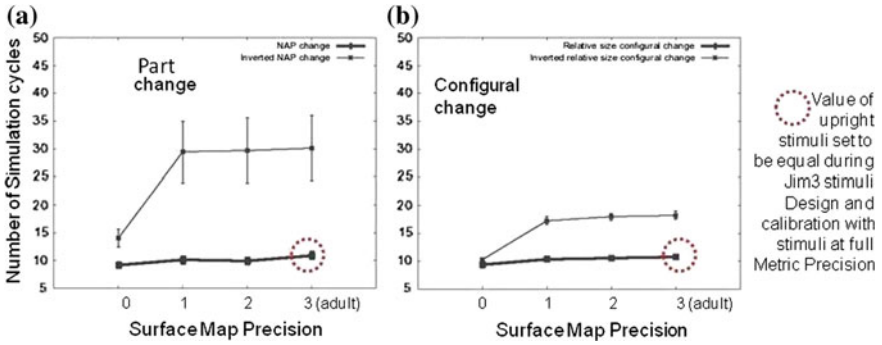

Fig. 11 Showing simulation results of animals and artefacts experiment with the surface map at four different levels of 'development'

with metric relation precision in the IGA decreased and Fig. 11 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 10 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 in [26].

Figure 11 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 from the empirical results presented in [26].

5.2.1 Statistical Analysis for Metric-Properties (MP) Manipulated Architecture

The simulation data were analysed with a 4 metric precision level: MP level 3 (adult with 45 neurons from three receptive field classes) versus MP level 2 (30 neurons from two receptive field classes) versus MP level 1 (15 neurons from one receptive field class) versus MP level 0 (no neurons) $\times 2$ (manipulation: part change versus relative size change) $\times 2$ (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) versus 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% CI: 2.18–3.82) 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) versus MP level 0 (no neurons). There was a non-significant difference in scores for adult MP level 3 (adult) (M = 3, 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% CI: −8.072 to 0.77) was large (eta squared = 0.85).

5.2.2 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 Sect. 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 per-

formance 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 metric 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 cycles to recognition (in the MP = 0 condition over 40 cycles).

The complication 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 from the adult performance level. This gave a very clear result. The relative size change condition showed significantly lower recognition performance for the youngest parameters, whereas the NAP change condition showed no significant difference between youngest and adult parameters, and a large effect size in the opposite direction to the relative size change condition (see Fig. 12). So the metric relation regressed simulations demonstrate a very clear dissociation in performance between the relative size change and NAP change conditions, just as the empirical results with human subjects show.

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

Fig. 12 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, $p = 0.054$)

5.2.3 Statistical Analysis for Surface-Map (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) versus SM level 2 (9 neurons in two further orientations from the centre) versus SM level 1 (5 neurons in one further orientation from the centre) versus SM level 0 (1 neuron with no further orientations from central neuron) $\times 2$ (manipulation: part change vs. relative size change) $\times 2$ (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) versus 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% CI: 0.36–2.11) 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) versus SM level 0 (1 neuron). There was a significant difference in scores for adult SM level 3 (adult) (M =3) 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% CI: 0.24–3.23) was moderate (eta squared = 0.047).

5.2.4 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 decreased. 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 highlight that the surface map regressed results do not match the empirical results reported in [26]. This test looked at whether the upright results for the 'youngest' SM regressed condition were significantly different from 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 from the empirical results reported by [26].

Figure 12 highlights the results of the post hoc *t*-tests for both the MP regressed and SM regressed architectures.

6 Conclusions

This paper shows that recent empirical results presented by Jüttner et al. [26] 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 by children in the 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 artificial 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 [26] 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 Figs. 7 and 10, 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 artefact and animal stimuli [26]. 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 experimental participants to eliminate distracters. In the JIM3 experiments time to recognition is always taken for stimuli presented on their own. The 'choose one from three 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 [14]. The simulations reported here present development just in terms of an increase in the numbers of neurons used for recognition in the IGA and the holistic surface map. So this explains changes in performance over development just in terms of differences in cognitive resources. We can also imagine an analogous theory of development from 'increasing metacognition' 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 [26]. 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 [26]. 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 centre 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 that differing life experiences and consequent encoding differences in memories might have on 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 suggest that children of age 7–9 years 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 such 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 adult 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 [26] experiment 3 and [25]). The finding that JIM3 needs to have no metric relation precision to qualitatively match 7–9-years-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.

In addition to just thinking about the four limitations noted above for how the computer simulation matches the task used in the human experiments, we can also consider that the human experiments are a limited approach in capturing the complexities of object recognition in more ecologically valid contexts. For instance, the human recognition task modelled in this paper involves a participant sitting passively whilst being presented with images, which do not move and cannot be acted upon or manipulated. This is partly done to conserve clear experimental control between the experimental conditions. However, it does have the downside of limiting possible mechanisms of active perception, such as the development of sensorimotor contingencies, and so limiting the role of active perception mechanisms which may not rely on explicit representations. Future work may involve more active experimental tasks, and modelling these observations with robots rather than disembodied simulations.

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 [27].

Secondly, 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 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 [16, 40]. 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. Alternatively, as Thoma and Henson's imaging results suggest, there may be two streams: dorsal and ventral, where the dorsal stream is involved in solely view-based recognition and the ventral stream involves some view-based as well as compositional recognition. Evidence for this complex arrangement is presented by Thoma and Henson, who noted that in their imaging studies: "*the ventral stream regions also showed greater RS from intact than split primes, which would not be expected if these regions utilised purely structural representations*" ([40], p. 524). This finding makes intuitive sense if we think of attention building up over the briefest of moments in time rather than starting 'on full'. So before attention can link object parts dynamically with relations, this system will already be decomposing parts and these isolated and independent parts may trigger backward activation from memory traces before fully compositional memories are matched to fully compositional perceptual representations.

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; performance on simple behavioural tasks, such as measuring view-dependence/invariance to object recognition of rotated images, can act as a poor surrogate for 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, using a variety of investigative methods, including priming experiments, neuropsychological studies, brain imaging and computational modelling can provide convergent evidence and an elaborated view of how neural systems can support representational diversity in humans, other animals and machines.

References

1. Abecassis, M., Sera, M.D., Yonas, A., Schwade, J.: What's in a shape? children represent shape variability differently than adults when naming objects. J. Exp. Child Psychol. **78**, 213–239 (2001)
2. Barenholtz, E., Tarr, M.J.: Reconsidering the role of structure in vision. In: Markman, A., Ross, B. (eds.) The Psychology of Learning and Motivation, vol. 47. Elsevier, Amsterdam (2006)
3. Barenholtz, E., Tarr, M.J.: Visual judgment of similarity across shape transformations: evidence for a compositional model of articulated objects. Acta Psychol. **128**, 331–338 (2008)
4. Biederman, I.: Recognition by components: a theory of human image understanding. Psychol. Rev. **94**, 115–147 (1987)
5. Biederman, I., Bar, M.: One shot viewpoint invariance in matching novel objects. Vis. Res. **39**, 2885–2889 (1999)
6. Biederman, I., Gerhardstein, P.C.: Recognising depth rotated objects: evidence and conditions for three-dimensional viewpoint invariance. J. Exp. Psychol. Hum. Percept. Perform. **19**, 1162–1182 (1993)
7. Biederman, I., Gerhardstein, P.C.: Viewpoint dependent mechanisms in visual object recognition: reply to Tarr and Bulthoff. J. Exp. Psychol. Hum. Percept. Perform. **21**, 1506–1514 (1995)
8. Bulthoff, H.H., Edelman, S.: Psychophysical support for a two-dimensional view interpolation theory of object recognition. Proc. Natl. Acad. Sci. U.S.A. **89**, 60–64 (1992)
9. Davidoff, J., Roberson, D.: A theory of the discovery and predication of relational concepts. J. Exp. Child Psychol. **85**, 217–234 (2002)
10. Davidoff, J., Warrington, E.K.: The bare bones of object recognition: implications from a case of object recognition impairment. Neuropsychologia **26**, 279–292 (1999)
11. Davidoff, J., Warrington, E.K.: A particular difficulty in discriminating between mirror images. Neuropsychologia **39**, 1022–1036 (2001)
12. Doumas, L., Holyoak, K., Hummel, J.: The problems of using associations to carry binding information. Behav. Brain Sci. **29**, 74–75 (2006)

13. Doumas, L., Hummel, J., Sandhofer, C.: A theory of the discovery and predication of relational concepts. Psychol. Rev. **115**, 1–43 (2008)
14. Flavell, J.H.: First discussant's comment: what is memory development the development of? Hum. Dev. **14**, 272–278 (1971)
15. Foster, D., Gilson, S.: Recognizing novel three-dimensional objects by summing signals from parts and views. Proc. R. Soc. Lond. B **269**, 1939–1947 (2002)
16. Goodale, M.A., Milner, A.D.: Separate visual pathways for perception and action. Trends Neurosci. **15**(1), 20–25 (1992)
17. Hayward, W.: After the viewpoint debate: where next in object recognition. Trends Cogn. Sci. **7**, 425–427 (2003)
18. Hayward, W., Tarr, M.: Testing conditions for view point invariance in object recognition. J. Exp. Psychol. Hum. Percept. Perform. **23**, 1511–1521 (1997)
19. Heinke, D., Humphreys, G.W.: Attention, spatial representation and visual neglect: simulating emergent attention and spatial memory in the selective attention for identification model (SAIM). Psychol. Rev. **110**, 29–87 (2003)
20. Hummel, J.: Where view-based theories of human object recognition break down: the role of structure in human shape perception. In: Dietrich, E., Markman, A. (eds.) Cognitive Dynamics: Conceptual Change in Humans and Machines, pp. 157–185. Lawrence Erlbaum (2000)
21. Hummel, J.: Complementary solutions to the binding problem in vision: implications for shape perception and object recognition. Vis. Cogn. **8**, 489–517 (2001)
22. Hummel, J., Biederman, I.: Dynamic binding in a neural network for shape recognition. Psychol. Rev. **99**, 480–517 (1992)
23. Hummel, J., Stankiewicz, B.J.B.J.: Categorical relations in shape perception. Spat. Vis. **10**, 201–236 (1996)
24. Juttner, M., Muller, A., Rentschler, I.: A developmental dissociation of view-dependent and view-invariant object recognition in adolescence. Behav. Brain Res. **175**, 420–424 (2006)
25. Juttner, M., Petters, D., Wakui, E., Davidoff, J.: Late development of metric part-relational processing in object recognition. J. Exp. Psychol. Hum. Percept. Perform. **40**, 1718–1734 (2014)
26. Juttner, M., Wakui, E., Petters, D., Kaur, S., Davidoff, J.: Developmental trajectories for part-based and configural object recognition in adolescence. Dev. Psychol. **49**(1), 161–176 (2013)
27. Leek, E.C., Reppa, I., Arguin, M.: The structure of three-dimensional object representations in human vision: evidence from whole-part matching. J. Exp. Psychol. Hum. Percept. Perform. **31**, 668–684 (2005)
28. Logan, G.D.: Spatial attention and the apprehension of spatial relations. J. Exp. Psychol. Hum. Percept. Perform. **20**, 1015–1036 (1994)
29. Marr, D., Nishihara, H.K.: Representation and recognition of the spatial organization of three-dimensional shapes. Proc. R. Soc. Lond. B **200**, 269–294 (1978)
30. Olshausen, B., Anderson, C., Van Essen, D.: A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. J. Neurosci. **57**, 4700–4719 (1993)
31. Poggio, T., Edelman, S.: A network that learns to recognize three-dimensional objects. Nature **343**, 263–266 (1990)
32. Rentschler, I., Juttner, M., Osman, E., Mller, A., Caelli, T.: Development of configural 3D object recognition. Behav. Brain Res. **149**, 107–111 (2004)
33. Stankiewicz, B.: Empirical evidence for independent dimensions in the visual representation of three-dimensional shape. J. Exp. Psychol. Hum. Percept. Perform. **28**, 913–932 (2002)
34. Stankiewicz, B.J., Hummel, J.: Automatic priming for translation- and scale-invariant representations of object shape. Vis. Cogn. **6**, 719–739 (2002)
35. Stankiewicz, B.J., Hummel, J., Cooper, J.E.: The role of attention in priming for left-right reflections of object images: evidence for a dual representation s of object shape. J. Exp. Psychol. Hum. Percept. Perform. **24**, 732–744 (1998)
36. Sutherland, N.S.: Outlines of a theory of visual pattern recognition in animals and man. Proc. R. Soc. Lond. B **171**, 95–103 (1968)

37. Tarr, M., Bulthoff, H.: Is human object recognition better described by geon-structure-descriptions or by multiple views? J. Exp. Psychol. Hum. Percept. Perform. **21**, 1494–1505 (1995)
38. Tarr, M., Pinker, S.: Mental rotation and orientation dependence in shape recognition. Cogn. Psychol. **21**, 233–282 (1989)
39. Thoma, V., Davidoff, J.: Object recognition: attention and dual routes. In: Osaka, I., Rentschler, I., Biederman, I. (eds.) Object Recognition, Attention and Action
40. Thoma, V., Henson, R.N.: Object representations in ventral and dorsal visual streams: fMRI repetition effects depend on attention and part-whole configuration. Neuroimage **57**, 513–525 (2011)
41. Thomai, V., Davidoff, J.: Priming of depth-rotated objects depends on attention and part changes. Exp. Psychol. **53**, 31–47 (2006)
42. Thomai, V., Davidoff, J., Hummel, J.: Priming of plane-rotated objects depends on attention and view familiarity. Vis. Cogn. **15**, 179–210 (2007)
43. Thomai, V., Hummel, J., Davidoff, J.: Evidence for holistic representations of ignored images and analytic representations of attended images. J. Exp. Psychol. Hum. Percept. Perform. **30**, 257–267 (2004)
44. Triesman, A.M., Gelade, G.: A feature-integration theory of attention. Cogn. Psychol. **12**, 97–136 (1980)
45. Ullman, S.: Object representations in ventral and dorsal visual streams: fMRI repetition effects depend on attention and part-whole configuration. Cognition **32**, 193–254 (1989)
46. Ullman, S.: Three-dimensional object recognition based on the combination of views. Cognition **67**, 21–44 (1998)
47. van der Velde, F., de Kamps, M.: Neural blackboard architectures of combinatorial structures in cognition. Behav. Brain Sci. **29**, 37–108 (2006)
48. Wolfe, J.M., Cave, K.R., Franzel, S.L.: Guided search: an alternative to the feature integration model for visual search. J. Exp. Psychol. Hum. Percept. Perform. **15**, 419–433 (1989)