KARMIC TABULA RASA

Karthik Mahesh Varadarajan
Technical University of Vienna, Vienna, Austria
kv@acin.tuwien.ac.at

Abstract

In this paper, we hypothesize a two-step process to visual perception and recognition of objects. The first level is based on an evolutionary cognitive algorithmic process which we label as 'karmic' (k). The second level depends on repeated learning of correlated local features in the object space, which we call the 'tabula rasa' (TR). We also present models for 'k' and 'TR' features grounded in our theory of Recognition by Component Affordances (RBCA). Using the combined features, which we call k-TRONs composed of 25 structural affordances and augmented by 10 material affordances, we support recognition of over 200 categories of commonly occurring objects. We argue that k-TRONs form the basis of evolutionary object recognition and present psychophysical and neurobiological evidence to support our hypothesis. Allied aspects of recognition such as Novelty detection, Equivalence classes, Recognition of articulated and natural objects, Attention, Saliency, Memory and Scale of analysis are also addressed.

Recognition by Components (RBC)

Biederman's *Recognition by Components* (RBC) puts forward the theory that only a modest number of geons (< 40) are essential in representing the wide variety of objects in the world. This theory represents the first holistic attempt at visual object perception. RBC also states that perception of these geons is viewpoint invariant, which is attributed to their being distinguished by three nonaccidental properties (NAPs) of contours which do not change with orientation in depth. However, RBC suffers from several serious drawbacks. A multitude of objects map to the same geon constellations. Furthermore, the theory uses only shape as a distinguishing factor, which makes classification between an orange and a tennis ball intractable. In a practical world, not all objects can be represented by geons. Articulated objects such as books, highly fractal objects such as shrubs and deformable objects such as clothing cannot be even roughly approximated by geons. Yet, humans have no difficulty in perceiving and recognizing these objects from a variety of poses and deformation states. Clearly, there are more complex mechanisms at work than simple recognition by geon configurations.

k-TR and RBCA

In this paper, we present an evolutionary model to visual perception and recognition of objects. This theory is grounded in our '*Recognition by Component Affordances*' (RBCA) model, which is an extension of RBC. We hypothesize that learning of object models in humans for recognition occurs in a two-step process. These steps are interdependent. The first or higher cognitive level of visual processing is based on an evolutionary cognitive algorithmic process which we label as '*karmic*' (k). The second level depends on reinforced learning of correlated local features in the object space, which we call the '*tabula rasa*' (TR) process. We hypothesize that the learning of object features for recognition in humans is the result of this combination of cognitive processes. '*k'* and 'TR' form dependent processes wherein, knowledge accumulated as part of 'TR' is transferred to the '*k'* process over long periods of time which performs implicit semantic deductive and inductive reasoning. This reasoning is based on functional affordances and form fits, as defined by our RBCA theory. These processes also parallel evolution of object models over the ages, for which it is possible to employ explicit semantic deductive and inductive reasoning. In the case of evolutionary

object representations using the '*k*' and 'TR' processes, the 'TR' representation of objects is essentially transient and varies with the constancy epoch – a period during which the 'TR' representation holds consistent for the object. On the other hand, the '*k*' component represents object constancy, i.e. the ability to recognize the object in a variety of poses, colors, lighting conditions etc. without having seen these instances previously and its description of an object holds consistent with minor incremental knowledge assimilation over several TR constancy epochs. In other words, the evolutionary '*k*' component is '*karmic*' and hold truth data about the visual world that can be consistent for several generations. 'TR' on the other hand forms temporal processes with truth data that are valid only for a short time (say a few years), during which time '*k*' derives knowledge from it if there are sufficient statistics of repeatability. The knowledge represented by 'TR' for a particular object can be re-instantiated or brought to a clean state if correlated instances of the object in question are not observed regularly, leading to '*tabula rasa*'. During this re-instantiation, old 'TR' records are moved to a deprecated store, which is reactivated only during an exhaustive object matching process. These two components together form '*k*-TR' or *Karmic Tabula Rasa*.

k-TRONs

We hypothesize that while 'TR' features are predominantly a multi-scale combination of (a) global features such as simple 2D shapes such as squares, circles or 3D shapes such as cylinders, cubes, (b) local features such as color or coarse texture and (c) semantic features such as co-occurrence likelihoods like that of pens and notebooks or spatial relationships such as books on top of tables etc. On the other hand, '*k*' features are essentially affordance features as defined by our RBCA theory. Each of the three different types of 'TR' features create corresponding '*k*' features. Any object recognition is the result of matching of all or a few of this set of six feature definitions. We call the combined features '*karmic Tabula Rasa ions*', which form the fundamental units of object representation and perception, in short, k-TRONs and argue that these form the basis of visual object recognition. The three affordance features ('*k*' features) corresponding to the three 'TR' features are labeled as (a) *Structural affordances,* (b) *Material affordances* and (c) *Semantic affordances*. Structural affordance is the primary and most crucial affordance for object definition and recognition. Structural affordance and Material affordance together form *Functional affordances* or *Functional form fits*. Semantic affordance on the other hand is an auxiliary affordance feature.

The Affordance Network

(a) *Structural affordance* corresponds to inferred knowledge about the 2D/3D shape of the object. For e.g., detection of a cylindrical or circular shape corresponding to a part of the object or as a whole, indicates a *Roll-ability affordance*. Example of an object affording *roll-ability* is the vehicle tire. Similarly, a flat or slightly convex part indicates a structural and hence function affordance of *Support-ability,* indicating that the part/object is capable of supporting other objects over it. Early definitions of affordances such as Gibson's *sittability* affordance can be mapped to this affordance. The seat of a chair is an example of an object part that provides the *support-ability* affordance. (b) *Material affordance* corresponds to deduced knowledge about the material properties of an object based on the local visual features. A good example is the shiny or grey color of a metallic object which results in an inference of high strength and durability of the object. This *Durability affordance* results in a knowledge inference stating that the object is *functionally fit* or offers a *functional affordance* that demands such durability. Another example of material and hence functional affordance stems from the object being colored black. These objects exhibit either *Socio-cultural preference conditioning* or *Environment conditioning* or *Stain suppression affordances*. Examples of such objects include electronic goods, winter wear and vehicle tires respectively.

(c) *Semantic affordance* is dependent on co-occurrence or typical pose of objects. Semantic affordances define the *Subject-Object* relationships for functional affordances. For e.g., the semantic affordance corresponding to the co-occurrence of pens and books can be defined based on the *Engrave-ability* affordance where the pen, book form a subject-object pair.

Table 1. The affordance network

				Part Affordance	Definition	Geometric Mapping	Examples
Structural	Intrinsic	Simple		Contain – ability	Holds solid or liquid transferred to it	High convexity	Empty bowl, Cup, Bag, Pot
				Support – ability	Supports objects in a horizontal position	Flat – Convex	Plate, Table, Chair, Stair, Plate, Spatula, Cabinet, Shelf
				Intrinsic contain –ability	Stores solid or liquid internally	Cylinder/Cube/Cuboid/Prism	Canister, Box, Barrel, Suitcase
				Incision – ability	Creates an incision or a cut	Sharp edge (flat linear surface)	Knife, Screwdriver, Boat hull, Airplane
				Engrave – ability	Creates a point depression	Sharp Tip	Cone, Pen, Drill
				2D Roll – ability	Rotates to move in along a linear path	Circular/ Cylindrical	Tire, Paper roll, Wheels, Fire extinguisher, Pipe, Gas cylinder, Submarine
				3D Roll – ability	Rotates to move at all angles	Spherical	Ball
				Contact – ability	Provides support for contact with another surface	Flat – Concave	Hammer head, Paddle
				Display – ability	Provides surface for display	Flat	Screens, Boards, Posters, Leaves, Paper, Camera
				Glide – ability	Glides in air in a horizontal position	Thin rectilinear	Aircraft wings, Rotor blades, Fins
				Flow-support – ability	Supports streamlined flow of liquids, solids or other	Circular/Cylindrical/Conic cavity	Pipes, Faucets, Funnel, Gun barrel, Hydrant, Camera, Flashlight, Flute, Clarinet, Turret
				Grab - support – ability	Provides surface with friction for precision grab - Relaxable to Stackability/ Grasp Affordance for imprecise cases like door handles	Hexagonal/Pentagonal Prismatic/Contour Cavity Texture	Pencil, Bolt, Gears
				Wrap - support – ability	Provides smooth surface for wrapping of other objects	Circular/Cylindrical	Roll, Pole
		Joint		Connect – ability	Connects to a corresponding mating socket	Solid with support (m)	Plug, USB Stick, Key, Bulb
		Texfordances	Internally Connected	Weed – ability	Weed through linear structures	Linear textural structures	Comb, Brush
				Filter – ability	Filter through linear structures	Bi-linear textural structures	Grid, Filters
				Connected Support – ability	Support objects between connected textural structures	2 or more connected Flat/Concave structures	File, Binder, Paper punch, Tongs, Pincers, Headphones, Eyeglasses,
			Externally Connected	Disconnected Support – ability	Support objects using an additional entity connecting textural structures	2 or more disconnected	Chopsticks, Columns, Pillars, Table legs, Wheels
	Extrinsic			Wrap(p) –ability	Serves to partially or wholly wrap around a predefined shape	w.r.t. given shape	Shoe, Glove, File, Spanner, Bottle opener, DVD case, Mouse, Umbrella, Toaster, Watch, Book, Guitar, Violin
				Compact – ability	Provides compact packing of multiple self-units	Cube/Cuboid	Box, Suitcase
				Stack - ability	Provides stable stacking of multiple self-units with respect to a pole	Ring/Torus/Toroid	Disks, Rings, Hoops, Donut
	Interfordances			Engrave + Incision - ability	Provides structure intermediate enabling both affordances	Enlarged sharp tip with edge	Axe
				Contact + Engrave – ability	Provides structure intermediate enabling both affordances	Enlarged tip with surface	Eraser
Material				Environment conditioning	Conditioning from reflective (white) and absorptive (black) requirements	White, Black	Chinaware, Walls, Sails, Ships, Summer/Winter wear
				Disposability	Disposable materials – paper/plastic	White	Forks, Papercups, Filters
				Elegance + Durability	Durable, ornate materials	Gold, Silver, Copper	Saxophone, Utensils, Sword
				Durability	Hard, durable materials	Grey, Silver	Trash can, Utensils,
				Stain suppression	Shaded to cloak dirt	Black	Tires, Telephones, Guns, Suitcase
				Socio-cultural preference conditioning	Colored based on cultural preferences – electronics	Black, Silver, White	Cell phones, Mouse, Watch
				Disposability + Durability	Durable, yet disposable materials – wood	Brown	Casket, Hammer, Table, Door, Comb
				Climatic Conditioning	Evolutionary colors for natural objects	Brown, White, Black	Animals
				Perceptivity	Shaded to enhance perceptivity	Red	Hydrant, Signboards, Fire extinguisher, First aid box
				Socio-cultural variant textures	Cultural text variations	Digits, Symbols, Characters	Clock, Calculator, Keyboard, Remote control

Again, it should be noted that the *Part Functional-Structural Affordance* is the most important for object recognition. Just as in the case of the geon theory, our study indicates that a restricted set of 25 structural part affordances together with our Part Joint Topological Relationship Schema [Varadarajan 2011-2], which are abstract labels for part topology and linkage, are sufficient in describing over 200 commonly occurring man-made object classes in everyday environment. This forms a major improvement over earlier affordance schemes such as Gibson's which explode in the dimensionality of the functions offered. The *Part Functional-Structural Affordance Schema* can be categorized into three affordance mechanisms. These are (a) *Intrinsic affordances* are defined entirely with respect to the object at hand, independent of its environment. These are further sub-divided into (i) *Simple affordances*, which are single part affordances such as Contain-ability, Support-ability, etc. (ii) *Joint affordances*, which are multi-heterogeneous part affordances (iii) *Repetitive or Textural affordances* – in short *Texfordances*, which are multi-homogeneous part affordances. These can be internally connected affordance networks or externally connected. (b) *Extrinsic*

affordances are defined with respect to other objects (c) *Interfordances* are intermediate simple affordances, while *(d)Polymorphic affordances* pertain to objects whose affordance state had been modified at some point in time. These are listed in detail in Table 1.

Novelty Detection and Equivalence Classes

Humans are capable of categorizing previously unseen objects with uncharacteristic color, texture, size and shape into its right class, without apriori knowledge about the exact instance name of the object. We hypothesize that this ability of humans is due to the inference of abstract knowledge about the object based on our theory of '*Conceptual Equivalence Classes*'. These classes are defined as sets of objects that are interchangeable from the view-point of usage for the primary functionality of the object. Hence, objects such as mugs, cups and beakers form an equivalence class. Bags and baskets also form an equivalence class, so do cans and bottles, bikes and motorbikes and so forth. All equivalence classes can be uniquely defined and recognized in terms of their (a) *Part Structural-Functional Affordance Schema* and (b) *Part Grasp Affordance Schema*. The *part grasp affordance schema* defines the structural grasp-ability of the object. We hypothesize that it is this recognition of the conceptual equivalence class of an object that enables humans to process and recognize the category of the wide variety of novel objects in the world, without being perturbed by socio-cultural design variations. Thus, the Structural Affordance schema forms core and primal mechanism for object constancy and recognition.

Recognition of Articulated and Natural Objects

The extrinsic affordances in our framework, such as *Wrap-ability* lend themselves to the modeling of articulated objects. For e.g. shoes can be modeled as wrap-able objects with respect to the model of feet, socks with respect to feet and the lower legs, stockings with respect to feet and legs, heels with respect to a tiptoe. Thus, all articulated objects in their natural state (with respect to the state of the afforded object), can be recognized using this framework. Surprisingly, the *k*-TR framework can be used for natural objects, albeit at a theoretical level. The distinctive features of animals through evolutionary processes in their environment or habitat can be explained in terms of affordance models. Elephants have evolved to have strong legs that provide the *texfordance* of support and trunk to afford *wrap-ability* of barks. Fish have fins that provide *glide-ability* in water and require no legs. Nevertheless, the extension of affordance models to such entities is rather difficult in comparison with man-made objects due to high levels cognitive reasoning required to deduce such affordances.

Attention, Saliency and Object Identity Retrieval

During recognition, '*k*' and 'TR' memory stores are searched for a match in the object feature space. We hypothesize here that object saliency detectors in the human visual perception system extract coarse grained 'TR' features, which activate all coarse grained 'TR' records (at coarse scales) that match approximately. These, in turn activate the fine grained 'TR' records, resulting in an iterative narrowing of object search space. The next stage is the activation of a search in the '*k*' records, which are responsible for the final object identity retrieval. This hypothesis is also in alignment with the reduction in recognition speeds for novel objects or objects in novel semantic contexts, as the there is very minimal object search space reduction in the 'TR' stages for novel objects and much of the actual recognition can be attributed to the '*k*' processing. In the event of a non-match, an exhaustive search using deprecated 'TR' records is carried out resulting in longer response times for recognition.

Scale of Analysis and Choice of Features

Scale of object analysis also plays a very important role in feature extraction and recognition. It is noted that humans do not always process objects at the finest scales for recognition. Using our RBCA theory, we hypothesize that there is a critical number of parts (5-6) beyond which humans do not traditionally decompose an object further. The remaining parts are

usually identified as material textures (especially when there are more than 2 repeated or identical parts).

Evaluation of the *k*-TR Hypothesis

The *k*-TR hypothesis is evaluated on five counts – (a) theoretical evaluation on a large database of 200 commonly found man-made objects (b) psychophysical priming evidence using a set of test images (c) neuro-biological grounding (d) evidence from language acquisition studies (e) computer vision models and algorithms.

(a) Theoretical Evaluation: k-TR is found to be valid and consistent in all cases of a sample database containing 200 everyday use man-made objects. Here, we demonstrate it using examples such as 'automobiles' and 'telephones'. All conventional automobiles provide '*2D Roll-ability*' or '*Linear Motion*' affordance through the functional form fit rendered by the cylindrical/ circular wheels, which form the '*k*' component, in addition to '*material affordances*'. All other design variations, such as commonly occurring colors, shapes, textures etc. are stored as 'TR' records that are valid only for the 'TR' epochs and have to be re-instantiated at the end of the epoch. This is more complicated in the case of 'telephones' wherein the '*k*' records hold only *socio-cultural variant textural affordances* in the form of digits on the numeric pad and no *structural affordances* can be perceived for object constancy. Hence features such as shape, color, size for the large variety of cell phones, land phones etc. are stored as 'TR' records which are re-instantiated every few years.

(b) Psychophysical Priming Evidence: In order to test the validity of our hypothesis, we conducted priming tests (subject to legality and ethical requirements) across 10 individuals of ages between 20-50 and measured response times and error rates for recognition. In these tests, the individuals were verbally unambiguously primed for a certain object label and were asked to register a positive response if an object corresponding to that label was observed in 5 sets of RSVPs of arbitrary images. The sets contained examples of the object as (a) simple sketch in the most commonly occuring shape (b) a real instance of the object in default color, texture and shape (c) real instance with default color, texture, shape and additionally context (d) real instance of the object but in an uncommon state (belonging to a different TR epoch than the current context) (e) novel instance of the object. These tests were conducted across 23 object category instances and the average response time and error rates were measured.

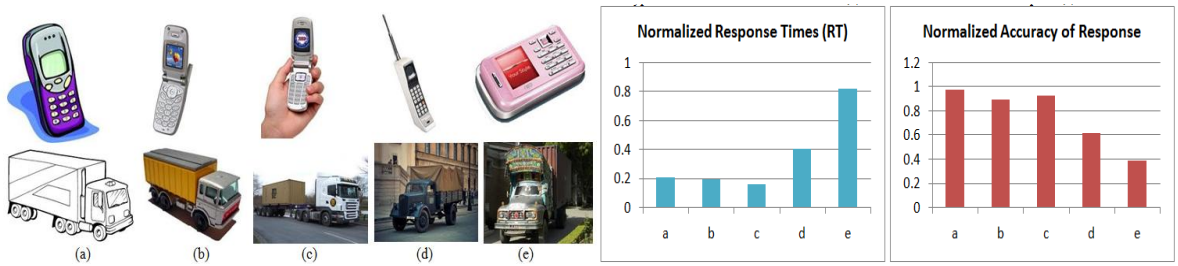

Figure 1. [Left] Sample images corresponding to 'cell-phone' and 'container truck' (a) sketch (b) typical instance (c) typical instance with semantic context (d) instance from a different TR epoch (e) novel instance [Right] Normalized average response times corresponding to (a)-(e) instances of object categories and Corresponding normalized accuracy rates

It can been observed that while material and semantic contexts improve recognition rates and accuracy (corresponding to the faster and efficient search for 'TR' features), the accuracy and speed of recognition drops significantly for instances from deprecated or empty 'TR' epochs and also for novel objects.

(c) Neuro-biological Grounding: It has been observed that the encoding of invariant visual primitives may relate to evolutionary and developmental plasticity mechanisms that shape object selectivity and invariance by taking into account statistical regularities in the environment. At a shorter time scale, experience through everyday dynamic interactions with

objects may cause tuning of neural populations in a probabilistic manner, result in invariance for feature configurations that co-occur frequently [Biederman 2009]. The *k*-TR hierarchy is also supported by the fact that while local image features are known to be processed in primary visual cortex, neural populations in higher temporal areas contain information about object constancy [Grill-Spector 2004, Hung 2005, Quiroga 2005]. Further neuro-biological evidence is necessary to support the choice of affordance features.

(d) Evidence from Language Acquisition Studies: At 18 months of age, children have a vocabulary of less than a 100 words [Siegler, 1986]. Since the vocabulary is insufficient to describe all a child wishes to, it results in overextension errors - erroneously extending the meaning of words in the existing lexicon to cover things and ideas for which a new word is lacking. For e.g., the general term for any kind of four-legged animal may be 'doggie', not just dogs. This choice of words is explained by the Feature hypothesis [Clark, 1973], which suggests that children form definitions that include too few features, such as anything with four legs is a 'doggie'. An alternative functional hypothesis [Nelson, 1973] suggests that children associate words with functions or purposes, resulting in overextension by function.

Significance of '*k*-TR' Hypothesis to Computer Vision Recognition Algorithms

It should be noted that our k-TR framework is conducive to direct implementation on computer vision software systems. This theory also presents the first holistic attempt to modeling and recognizing objects for computational visual perception. A first attempt at modeling the '*k*' component of '*k*-TR' using RBCA theory is presented in [Varadarajan 2011-1]. A practical approach to knowledge assimilation and representation for recognition of Equivalence Classes and objects is demonstrated in our work in [Varadarajan 2011-2]. A part based approach to recognition using 'TR' features has been demonstrated in [Varadarajan3 2011-3]. This paper also expounds a first approach towards implementing a holistic '*k*-TR' system using '*k*-TRON' features. Actual implementation of the entire multiple ontology framework (the most important ontology being our affordance network) is ongoing work.

References

Siegler, R. S. (1986). Children's thinking. Englewood Cliffs, NJ: Prentice-Hall.

Clark, E. V. (1973). What's in a word? On the child's acquisition of semantics in his first language. In T E. Moore, Cognitive development and the acquisition of language. New York: Academic Press.

Nelson, K. (1973). Structure and strategy in learning to talk. Monograph of the Society for Research in Child Development, 38(Serial No. 149).

Biederman I, (1987) "Recognition-by-components: a theory of human image understanding" Psychological Review 94 115-147.

Hung C P, Kreiman G, Poggio T, DiCarlo J J, (2005), "Fast readout of object identity from macaque inferior temporal cortex", Science 310 863-866.

Quiroga R Q, Reddy L, Kreiman G, Koch C, Fried I, (2005), "Invariant visual representation by single neurons in the human brain", Nature 435 1102-1107.

Biederman, I., & Cooper, E.E. (2009), "Translational and reflectional priming invariance: A retrospective", Perception, 38, 809-825.

James J. Gibson (1977), "The Theory of Affordances", In Perceiving, Acting, and Knowing, Eds. Robert Shaw and John Bransford, ISBN 0-470-99014-7.

KM. Varadarajan, M. Vincze, (2011-1) "Object Part Segmentation and Classification in Range Images for Grasping", IEEE International Conference on Advanced Robotics – ICAR.

KM. Varadarajan, M. Vincze, (2011-2), "Knowledge Representation and Inference for Grasp Affordances", International Conference on Computer Vision Systems – ICVS.

KM. Varadarajan, M. Vincze, (2011-3) "Holisitic Visual Cognitive Recognizer using Part based Local, Global, Semantic and Affordance Features", IEEE CVPR FGVC.