BioSystems 135 (2015) 35–49

Contents lists available at ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems

Proto-object categorisation and local gist vision using low-level spatial features

Jaime A. Martins[a,c,*], J.M.F. Rodrigues[b,c], J.M.H. du Buf[a,c]

[a] *Vision Laboratory (FCT), LARSyS, Portugal*
[b] *Vision Laboratory (ISE), LARSyS, Portugal*
[c] *University of the Algarve, Portugal*

ARTICLE INFO

Article history:
Received 1 May 2014
Received in revised form 2 July 2015
Accepted 2 July 2015
Available online 14 July 2015

Keywords:
Disparity
3D
Stereo vision
Colour
Population coding
Learning
Biological model
Figure-ground
Segregation
Object
Categorisation
Verification
Neural network
Visual cortex

ABSTRACT

Object categorisation is a research area with significant challenges, especially in conditions with bad lighting, occlusions, different poses and similar objects. This makes systems that rely on precise information unable to perform efficiently, like a robotic arm that needs to know which objects it can reach. We propose a biologically inspired object detection and categorisation framework that relies on robust low-level object shape. Using only edge conspicuity and disparity features for scene figure-ground segregation and object categorisation, a trained neural network classifier can quickly categorise broad object families and consequently bootstrap a low-level scene gist system. We argue that similar processing is possibly located in the parietal pathway leading to the LIP cortex and, via areas V5/MT and MST, providing useful information to the superior colliculus for eye and head control.

1. Introduction

There are many visual pathways related to object categorisation and recognition, especially focusing on quick shape categorisation, which is essential for scene gist. Obvious information sources for object shape and segregation are colour, texture, motion and depth from stereo. However, much richer information is available. Apart from depth from stereo, velocity gradients of optical flow can be used to locally encode ordinal depth at surface borders and also, but globally, ego-motion (Raudies et al., 2013). In addition, in case of occlusions, figure-ground segregation by local (and often intricate) border-ownership relations of e.g. vertex (keypoint) structures is also possible, as hypothesised in the discussion paper by Kogo and Wagemans (2013), which attracted many comments. How our visual system extracts and integrates all information is still rather speculative.

In this paper we focus on the transition between low-level syntax and low-level semantics, using elementary information such as surface lighting, colour and stereo disparity. The goal is to develop an integrated system for fast local gist vision: *which types* of objects are *about where* in a scene. This is necessary to bootstrap and guide, even alleviate, the processing in the ventral and dorsal data streams. These streams are known to serve two goals: the dorsal stream, also called the *where* or *vision-for-action* stream, is mostly devoted to optical flow and stereo disparity, whereas the ventral stream, also called the *what* or *vision-for-perception* stream, is devoted to object categorisation and recognition (Konen and Kastner, 2008; Farivar, 2009). However, the dorsal stream can also play a very important role in fast object categorisation (Gottlieb, 2007; Janssen et al., 2008; Konen and Kastner, 2008), which is the main focus of this paper.

[*] Corresponding author. Tel.: +351 2898001007751.
E-mail address: jamartins@ualg.pt (J.A. Martins).

http://dx.doi.org/10.1016/j.biosystems.2015.07.001
0303-2647/© 2015 Elsevier Ireland Ltd. All rights reserved.

Nomenclature

AIT	anterior inferotemporal cortex
AUC	area under ROC
CMC	cumulative match characteristic
DEM	disparity-energy model
DET	detection error trade-off
DoG	difference-of-Gaussians
EER	equal error rate
FAR	false acceptance rate
FoA	focus-of-attention
FRR	false rejection rate
HTER	half total error rate
IT	inferior temporal cortex
LGN	lateral geniculate nucleus
LIP	lateral intraparietal cortex
MST	medial superior temporal cortex
PIT	posterior inferior temporal cortex
RF	receptive field
ROC	receiver operating characteristic
ROR	rank-one recognition
SC	superior colliculus
SVM	support vector machine
V1	primary (striate) visual cortex
V2	secondary (prestriate) visual cortex
V4	visual area V4 of the extrastriate visual cortex
V5/MT	visual area V5 of the middle temporal visual cortex

An integrated system must first solve two hard problems: (a) the first one is of paradoxical nature, as precise object categorisation and recognition in the ventral stream requires object segregation, but object segregation has usually been regarded only possible if the system already knows what the object is (assuming of course that objects are complex and that they are seen against equally complex backgrounds). Consequently, we explore the possibility for a system to segregate an image region or an object even before knowing what it is, based on robust low-level and local shape features like edge conspicuity and disparity. (b) The second problem is that object categorisation and recognition is a sequential process: while fixating one object, its features must be routed to normalised object templates held in memory. This routing blocks the system until categorisation and recognition have been achieved, after which the system is released for dealing with another object. Therefore Rensink (2000) proposed the concept of *proto-object* tied to a non-attentional "scene schema," consisting of concurrent spatial-layout and gist subsystems which both drive attentional object categorisation and recognition, all employing proto-object shapes resulting from low-level vision. Gist vision addressed so far mostly concerns global gist of entire scenes (Bar, 2004; Siagian and Itti, 2007; Ross and Oliva, 2010; Rodrigues and du Buf, 2011). However, there is already some research into local parts of scenes that allow a quick pre-categorisation of geometrically shaped objects (Martins et al., 2012). Here we extend the use of fixed geometric shapes and aim at representing any kind of proto-object shape.

The use of depth information has shown good results in the context of general object detection (Quigley et al., 2009), as depth information is resilient to lighting and tonal variations, provides geometrical hints and is efficient for separating foreground from background. Most recent categorisation and recognition methods use RGB-D images and employ sophisticated features such as spin images for 3D point clouds (Johnson and Hebert, 1998), SIFT for 2D images (Lai et al., 2011), neural networks with deep learning (Socher et al., 2012), specific colour, shape or geometry features (Bo et al., 2011; Lu and Rasmussen, 2012) or even log-Gabor PCA (Gopalakrishna et al., 2014). Previous work on low-level shape feature extraction, detailed in Martins et al. (2012), used adaptive feature detectors for extracting corners, edges (bars) and curvature information from objects and defined simple but efficient rules for classification based on geometric relationships between those features. It was shown that many man-made objects with geometric properties obey those rules, yielding good results. We now propose a more elaborate and generic shape extraction method that can define a shape feature vector without explicitly constricting the feature-search space, generalising the process of object categorisation, also without needing to define geometric relationships between features. This method is then used for categorising 300 different objects of the RGB-D Object Dataset (Lai et al., 2011) (explained below) with 51 object classes.

Our main contribution is a biologically inspired framework for quick object detection and categorisation. It relies on salient scene information to simultaneously detect foreground objects, retrieve object shapes and disregard superfluous information. Proto-objects are built using normalised shapes, resulting from low-level attentional edge conspicuity and disparity processes. Edge conspicuity is a measure for object salience that highlights the transitions in colour/lighting at the borders of objects. It relies heavily on simultaneous colour and luminance contrast of an object with its background, so it is able to represent both salience and shape information. When combined with disparity information, there is often sufficient evidence for object detection, inhibiting the background of scenes and highlighting conspicuous objects in the foreground, so that robust object shape feature vectors can be obtained. These can then be used in a feed-forward processing scheme, like a neural network, to quickly assess a shape category, effectively being a type of "proto-object" representation that only needs few data points. This is especially useful for constrained processing systems that have limited resources and must first prioritise image regions or shapes to process, which is common in robotics. Apart from conspicuity, disparity information is also of paramount importance for shape categorisation and recognition, since it is only mildly affected by variations in pose and illumination. We also aim to prove that structural object information that is available from biologically inspired salience methods can successfully be applied to recognise objects with good accuracy, and is also capable of yielding information unavailable in luminance-based methods.

Section 2 explains the used databases, Section 3 details the steps necessary for the detection of objects and encoding of their shapes, Section 4 explains the experimental conditions and categorisation processes, Section 5 presents the evaluation trials and Section 6 the discussion and conclusions based on the data.

2. Object categorisation databases

An ideal database for the present research would need to fulfil four conditions: (I) Have RGB camera-rectified stereo image pairs of objects with (II) full object revolutions, (III) supply a large collection of objects with different categories and (IV) be actively used by several authors for object categorisation research, so that results can be compared and validated. Unfortunately, at present time, there is no database that complies to all, so we needed to use two databases, each for a different purpose.

2.1. CSCLAB Image Database

The CSCLAB Image Database (CSCLAB ID) (Murphy-Chutorian and Triesch, 2005) is one of the few to satisfy condition I but unfortunately not II–IV. It was created at the Complex Systems and Cognition Laboratory at the University of California, San Diego.

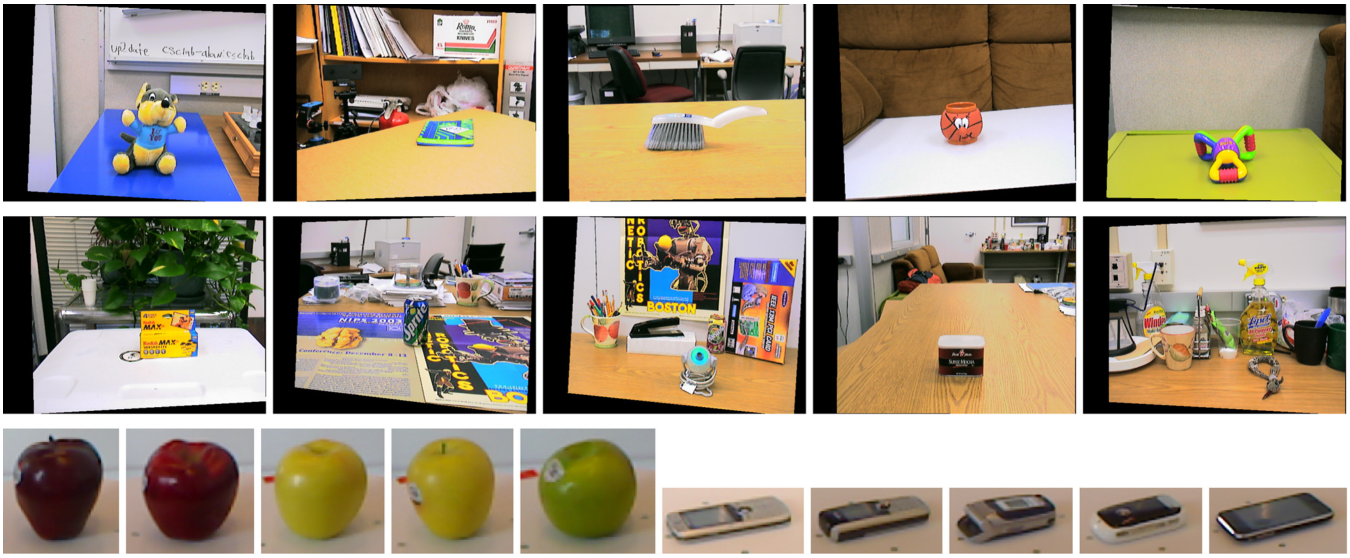


Fig. 1. Examples of objects. Top: CSCLAB image database; left stereogram images of 10 example objects with each of the 10 different backgrounds (first row: *blue table*, *bookshelf*, *computer desk*, *couch* and *green table*; second row: *plants*, *poster table*, *robot poster*, *table* and *tea table*). Bottom: RGB-D object database; five apples and five cellphones at 0° turntable position and 30° camera angle.

It consists of a single view of 50 mundane objects for training and 498 heterogeneous scenes for testing, each containing from 3 to 7 objects, with similar poses in 10 different backgrounds. Objects can be significantly occluded and display subtle differences in scale, viewpoint and illumination conditions. Data consists of RGB camera-rectified stereograms of objects in frontal views. This database is used here to illustrate the first part of our work, dealing with foregound/background segregation, object detection, segmentation and shape extraction, on different kinds of complex scenes/backgrounds (Section 3).

2.2. RGB-D Object Dataset

The RGB-D Object Dataset (RGB-D OD) (Lai et al., 2011) satisfies conditions II–IV but not I. It contains visual and depth images of 300 distinct objects of 51 categories, with many views, chosen from those commonly found in home and office environments where personal robots are expected to operate. Objects are organised into a hierarchy taken from WordNet hypernym/hyponym relations, which is a subset of the categories in ImageNet. Data was recorded with the cameras mounted at three different angles relative to a turntable where the object was located, at angles of approximately 30°, 45° and 60° with the horizontal plane. One revolution of each object was recorded at each angle. Each video sequence was recorded at 20 frames per second and contains about 250 frames, for a total of 250,000 RGB + Depth frames counting all objects. Unfortunately, this dataset only contains single RGB images (no stereo pairs), but with matching depth images for each object. Hence, it cannot be used for obtaining stereo disparities from the RGB images, forcing the use of the supplied depth images. We use this dataset to compare our object categorisation performances with those obtained by others. Since RGB-D OD was created by using a high-resolution RGB camera and an IR light pattern to measure disparity (a prototype RGB-D PrimeSense camera, similar to Microsoft's Kinect), it allows us to establish a baseline for expected performance of a system which employs precise depth information.

Examples of used object data are shown in Fig. 1. The top two rows show CSCLAB ID stereo-rectified images of objects in all 10 different backgrounds (only left viewpoint images are shown). We use the stereo image pairs for exemplifying our whole object processing algorithm, from detection to shape extraction, using disparity and conspicuity mechanisms. RGB-D OD contains cropped object images with corresponding range maps. Fig. 1 (bottom row) shows five example objects of categories "apple" and "cellphone." We use both range maps and RGB data to detect and extract object shape information and then estimate categorisation performance.

3. Object detection and shape coding framework

This section describes how low-level features can be combined to detect an object in a complex scene, yielding a binary segmentation mask of the object's outline. This mask is then used to extract a shape feature vector that describes the object's contour which is independent of object size, but dependent on object perspective. We will illustrate this process using images from the CSCLAB Image Database (Murphy-Chutorian and Triesch, 2005), that has each object in 10 different backgrounds. Below we introduce our biological disparity model, the edge conspicuity model (border saliency), the classifiers used and the classification rules in case of identification and verification experiments. For having a completely biological framework we will also consider – apart from the two disparity cell populations used, one for encoding and another for decoding – a third population: a neural-network classifier.

3.1. Disparity-based background inhibition

First, we apply a disparity energy model (DEM) implementation optimised for real-world images. It extracts disparity maps from all stereograms in the CSCLAB database. Since a detailed explanation is beyond the scope of this paper, we refer the reader to Martins et al. (2015, 2011), where this DEM implementation is thoroughly explained. Obtained disparities $D(x, y)$ for CSCLAB object number 107, a beer bottle (Fig. 2a), are exemplified in Fig. 2(b).

This disparity is then used for a foreground/background segregation, based on statistical properties of the scene. First, we discard areas where wrong disparity estimates are most commonly found: the top and bottom strips of the disparity maps (each strip with 1/6th of the vertical image size v), along with the left strip (1/8th of

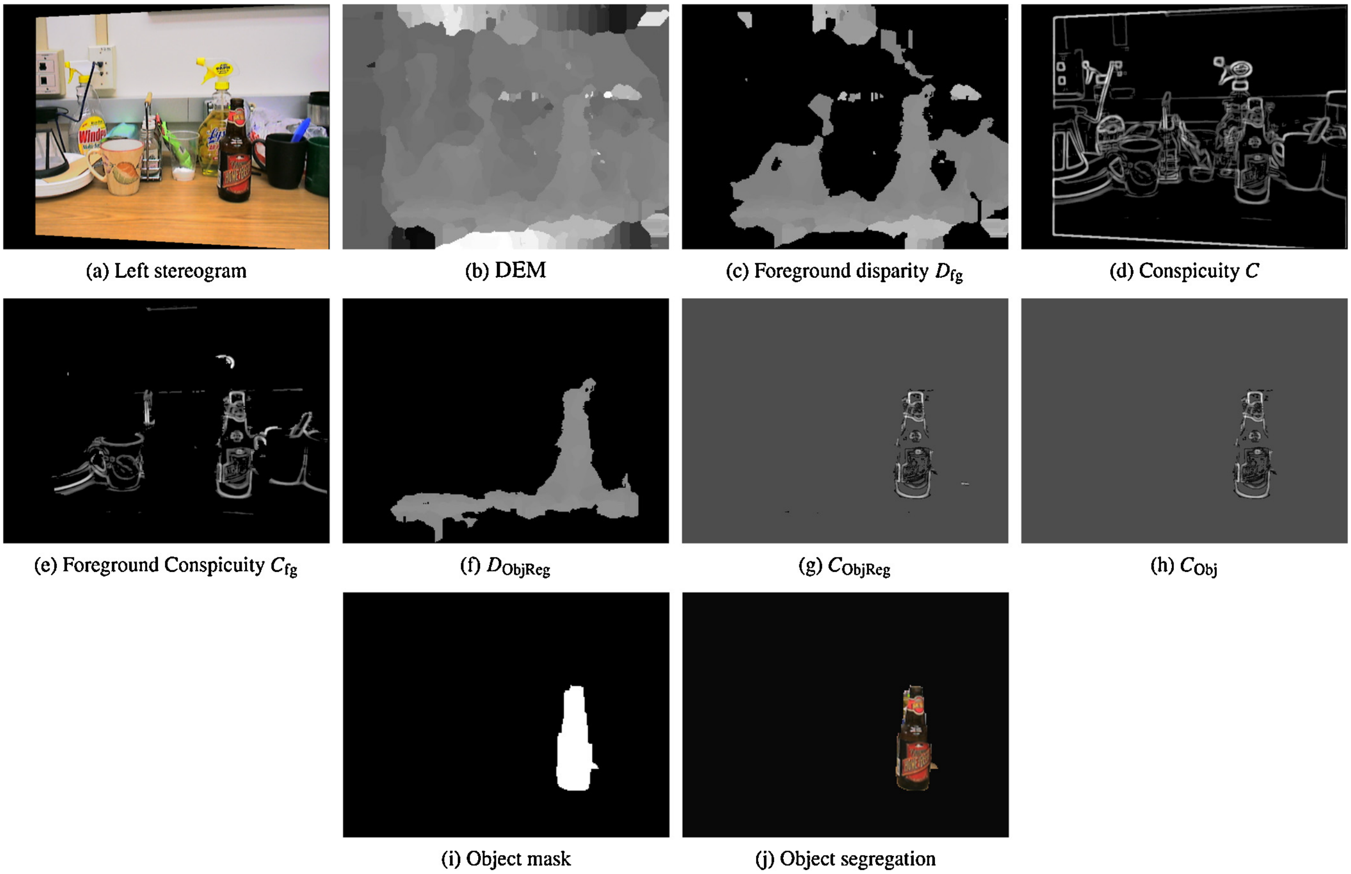

(a) Left stereogram
(b) DEM
(c) Foreground disparity D_{fg}
(d) Conspicuity C
(e) Foreground Conspicuity C_{fg}
(f) D_{ObjReg}
(g) C_{ObjReg}
(h) C_{Obj}
(i) Object mask
(j) Object segregation

Fig. 2. Beer bottle on *tea table* shape extraction. (a) Left image of the stereo pair; (b) DEM disparity map; (c) foreground disparity map; (d) scene edge conspicuity; (e) foreground scene conspicuity map; (f) foreground object region disparity; (g) foreground object region conspicuity; (h) object conspicuity; (i) object shape mask; (j) object segregation using the shape mask applied to the left stereogram image.

the horizontal image size h), resulting in the selected area $S(x, y)$, defined $\forall x, y$ as:

$$S(x,y) \triangleq \begin{cases} D(x,y), & \text{if } \ (x > h/8) \ \wedge \ (v/6 < y < 5v/6) \\ \text{OFF}, & \text{otherwise.} \end{cases}$$

From $S(x, y)$, we assign as foreground all disparities higher than the mean plus 0.1 times the standard deviation[1] and lesser than the mean plus two times the standard deviation (wrong estimates). The remaining disparity values are discarded. Mathematically, for an image of size $h \times v$, this results in a foreground – only disparity image D_{fg}, defined as:

$$D_{\text{fg}}(x,y) \triangleq \begin{cases} S(x,y), & \text{if } \ \bar{S} + 0.1\sigma_S < S(x,y) < \bar{S} + 2\sigma_S \\ \text{OFF}, & \text{otherwise.} \end{cases}$$

The notation $(\bar{\cdot})$ reflects the mean and $\sigma_{(\cdot)}$ the standard deviation. The OFF state represents discarded image data, which will either be the background or wrong disparity estimates. Overall, this step serves effectively as a disparity-based global scene pre-segmentation process that allows for posterior processing of only foreground regions. An example result can be seen in Fig. 2(c).

3.2. Edge conspicuity model

Edge conspicuity has been shown to yield good results in object shape discrimination, using luminance and colour differences to differentiate object shapes (Martins et al., 2009, 2012).

Succinctly, this model starts with adaptive colour smoothing, removing redundant information which is not necessary for shape detection, while preserving any local boundaries. This helps to stabilise differences between the inside and outside of objects. This is done using a cell layer that outputs a result similar to an adaptive difference-of-Gaussians (DoG) filter with edge preservation (for an in-depth discussion, please refer to Martins et al. (2012)). The resulting colour image from this step is defined as $I(x, y)$.

Considering the above step as a low-level process in the primary visual cortex (area V1), we encode $I(x, y)$ in CIELUV colour space[2] as $I_\alpha(x, y)$, with $\alpha \in \{L, u, v\}$. Conspicuity $\widetilde{C}$ is then defined as an edge salience measure that represents the maximum difference between colour triplets in $I_\alpha(x, y)$, at four pairs of symmetric positions from point (x, y), i.e., on horizontal, vertical and two diagonal lines (Martins et al., 2012). Conspicuity $\widetilde{C}(x, y)$ is defined as the maximum Euclidean distance of all four pairs,

$$\widetilde{C}(x,y) \triangleq \max_{i=1}^{4} \sqrt{\sum_{\alpha} [I_\alpha((x,y)-\boldsymbol{\delta}_i) - I_\alpha((x,y)+\boldsymbol{\delta}_i)]^2}.$$

The (x, y) coordinates of vector $\boldsymbol{\delta}_i$ are $(1, 0)$, $(1, 1)$, $(0, 1)$ and $(-1, 1)$.

[1] The 0.1 threshold was chosen as a good compromise for most cases. This is compatible with a scenario where a robot has to detect objects within a certain range, relying on calibrated stereo cameras for disparity extraction.

[2] CIELUV is a colour space built to attain perceptual uniformity of colour representations and mimics the double-opponent colour cells found in human vision, making it very useful for estimating the perceptual differences between image regions.

Only responses higher than 10% of $\max(\widetilde{C})$ are kept, in order to remove low-activity responses due to noise and also gradients caused by non-uniform illumination. This yields conspicuity edge positions

$$C(x,y) \triangleq \begin{cases} \widetilde{C}(x,y), & \text{if} \quad \widetilde{C}(x,y) > 0.1 \cdot \max(\widetilde{C}) \\ \text{OFF}, & \text{otherwise;} \end{cases}$$

see Fig. 2(d).

3.3. Foreground object detection

At this stage, we use a combination of conspicuity and disparity information to estimate possible foreground object locations. For the present work, we are only interested in processing the most conspicuous foreground object in each scene, i.e., the most salient, so other possible object locations are discarded. However, future research should address multiple salient objects for local gist vision, still as a fast and parallel process. This process could bootstrap another, but sequential, process in which overt attention by focus-of-attention (FoA) can scrutinise different regions for precise object recognition. Here, foreground objects are detected in two steps:

Foreground conspicuity. Since disparity and conspicuity extraction are done in parallel by two different cell populations, we also envision a quick V1/V2 low-level process that combines them for signalling important regions for low-level attention (Rensink, 2000) and FoA gaze. Since the D_{fg} map has all foreground pixel positions, the map

$$C_{fg}(x,y) \triangleq \begin{cases} C(x,y), & \text{if} \quad D_{fg}(x,y) \neq \text{OFF} \\ \text{OFF}, & \text{otherwise,} \end{cases}$$

will contain only the foreground conspicuity values. This is exemplified in Fig. 2.

Object region detection. The goal of this step is to detect the closest and most conspicuous regions in a scene, where possible objects can be located. It combines information from both foreground disparity and conspicuity maps. First, we count the number of active foreground conspicuity cells for each possible disparity 3-plane slice[3] $(d-1, >d, >d+1)$, with $D_{fg}^{\min} \leq d \leq D_{fg}^{\max}$. This can be defined as

$$A^C(d) \triangleq \sum_{i=d-1}^{d+1} \sum_{x,y} \{ \left[C_{fg}(x,y) \neq OFF \right] \quad \wedge \quad \left[D_{fg}(x,y) = i \right] \},$$

where $[\cdot]$ denotes a (binary) cell set.

Next, we select a single disparity plane slice δ using two simultaneous conditions: it is the closest possible disparity plane (highest d) with the highest cell count (highest $A^C(d)$ value for $\delta = d$). For this we use a weighted criterion, such that δ satisfies

$$\delta \triangleq \arg\max_d (A^C(d) \cdot d^4) : D_{fg}^{\min} \leq d \leq D_{fg}^{\max}.$$

The disparity value δ therefore represents the closest disparity plane slice that simultaneously has the most active conspicuity cells, and is also the best candidate for the closest foreground object. Since we wish to expand the disparity slice to encompass all possible disparity values of the object in question, as objects rarely occupy only a single disparity plane, we define a range parameter r such that the final disparity object slice will be in $\left[\delta - r, \delta + r\right]$, with $r \triangleq \sigma_S/2.5$ (the parameter 2.5 was empirically chosen[4]). The region of foreground disparity ranges where the object is located is then defined as

$$D_{ObjReg}(x,y) \triangleq \begin{cases} D_{fg}(x,y), & \text{if} \quad \delta - r \leq D_{fg}(x,y) \leq \delta + r \\ \text{OFF}, & \text{otherwise.} \end{cases}$$

This is illustrated in Fig. 2(f). We can see that only disparity values corresponding to the bottle's range are preserved, while the rest is discarded.

Object detection. The next object detection process selects only the conspicuity values of the foreground object, within the previous region. We determine C_{ObjReg}, which represents the significantly active conspicuity cells inside the object region, i.e., active-above-average, by

$$\begin{aligned} \psi(x,y) &\triangleq \begin{cases} C(x,y), & \text{if} \quad D_{ObjReg}(x,y) \neq \text{OFF} \\ \text{OFF}, & \text{otherwise;} \end{cases} \\ C_{ObjReg}(x,y) &\triangleq \begin{cases} \psi(x,y), & \text{if} \quad \psi(x,y) > \bar{\psi} \\ \text{OFF}, & \text{otherwise.} \end{cases} \end{aligned}$$

The result of this step is shown in Fig. 2(g): pixels darker than the background (which corresponds to the average $\bar{\psi}$) represent the less-than-average conspicuity values and are discarded, while pixels brighter than the background represent the most active conspicuity cells in the object's region and are kept in the C_{ObjReg} map.

Border refinement. There is now a further refinement step to see if all active conspicuity cells in the object's region effectively belong to a single object or if they can also correspond to parts of nearby objects, which should be discarded. This is done by keeping only the biggest connected area (i.e., a single object) in a four times morphologically disk-dilated (radius one) binary map where $\psi(x,y) \neq OFF$, which closes small gaps between active conspicuity cells. All pixels within this area in C_{ObjReg} are selected to form a single object conspicuity image C_{Obj}. The result is shown in Fig. 2(h), where a small dash to the right of the bottle was eliminated. This process can also be explained biologically using an equivalent higher-level grouping cell population with big receptive fields (RFs), for example in cortical area V2, which only activates when the lower C_{ObjReg} layer has enough active cells within each higher-level RF region.

3.4. Object mask

The next step uses the C_{Obj} cells, which are now segregated from everything else, to extract an object mask, which is useful for both segmentation and shape categorisation. This is done in two steps:

Non-maximum suppression. A cell layer C_{Obj}^M is built on top of the C_{Obj} layer. It applies non-maximum suppression in order to extract the positions where C_{Obj} has a local maximum in horizontal, vertical and diagonal directions, in 3 × 3 cell neighbourhoods. This is achieved by four oriented cell clusters plus one grouping cell at the output. For details, see Martins et al. (2012).

Contour continuity and filling. Contour gaps of C_{Obj}^M are closed using a process similar to morphological closing (four binary dilations followed by four erosions, using a radius-1 disk as structuring element). This results in F_{Obj}. The aim is to get a closed object contour, so that F_{Obj} can be used for segregation. The inside of the shape is then filled, yielding a binary segmentation mask. If the object's contour is still unconnected, there is only a partial fill (or none), depending on whether there were inside contours and those could be filled. All closed contours are filled, even those spaced

[3] or 2-plane slice for the lower and upper limits, $(d, >d+1)$ and $(d-1, >d)$ respectively.

[4] This value is proportional to the expected depth range of objects and is used to prevent clipping of object details that are not at the object's main disparity δ.

apart. In this case, the areas of all are calculated and those with an area less than 1/5th of the maximum are discarded. This allows for, at least, a partial object mask to be kept. An example mask is shown in Fig. 2(i). When applied to image (a) it results in image (j), with just the segregated object.

Fig. 3 shows results of the beer bottle shape in the other nine different backgrounds of CSCLAB ID. In only one case (poster table) the pattern on the table close to the bottle is so complex that segregation using only disparity and conspicuity information is not sufficient.

3.5. Shape feature vectors

Humans can recognise objects using several vision cues, even when seeing different views of the same object. We know that shapes of objects depend on the observer's perspective: the shape of a particular object can be rather stable when the observer's viewpoint rotates around it, e.g., an orange, or it can change significantly, e.g., a statue. Also, objects can appear with any arbitrary rotation, like upside-down or a bottle lying flat on a table, which is also a complication that must be solved.

For the scope of this paper, we chose to address the problem of observer perspective using canonical object poses, since it is the most relevant case for local gist processing, probably even hard-coded in a very quick proto-object neural pathway (Yanulevskaya et al., 2013; Martin and von der Heydt, 2013). It also makes sense from an evolutionary, survival perspective – predators in upright or running poses are much more dangerous than when they are lying down or even upside-down. Our approach combines the information to solve the shape-encoding problem using a hypothetical proto-object shape feature vector that represents shape from a common-oriented perspective viewpoint.

For implementation, our shape representation builds upon the centroidal-profiles methodology (Davies, 2004), which represents shape boundary distances in a polar coordinate system. For retrieving a shape vector $\boldsymbol{s}$ we first calculate the object's centroid coordinates. This is then assigned position (0, 0). The perimeter of the object is then followed counter-clockwise, from -180° to 180°. The retrieved perimeter positions are then converted to polar coordinates, yielding a rotation angle $\theta \in [-180^\circ, 180^\circ]$ and a corresponding distance ρ_θ. The retrieved values are sampled within a 1° interval, between $[\theta, \theta + 1^\circ]$, storing for each interval the $\rho^{\max}_{[\theta,\theta+1^\circ]}$ and $\rho^{\min}_{[\theta,\theta+1^\circ]}$ values. The first describes the outer-shape of the object and the second the inner-shape. Since general objects are not always star-shaped (i.e., with only a single intersection for each angle), both are needed to characterise objects and to make shapes immune to outliers. We then define vector $\boldsymbol{s}$ as

$$\boldsymbol{s} \triangleq \left(\rho^{\max}_{[\theta,\theta+1^\circ]}\Big|^{\theta=179^\circ}_{\theta=-180^\circ}, \rho^{\min}_{[\theta,\theta+1^\circ]}\Big|^{\theta=179^\circ}_{\theta=-180^\circ} \right).$$

This is a vector of 720 elements, 360 from $\rho^{\max}$ and 360 from $\rho^{\min}$. For the final feature vector values, we normalise $\boldsymbol{s}$ to be invariant to specific object sizes by subtracting the mean $\bar{s}$ and dividing by the standard deviation $\sigma_{\boldsymbol{s}}$,

$$\boldsymbol{s}^N \triangleq \frac{\boldsymbol{s} - \bar{s}}{\sigma_{\boldsymbol{s}}}.$$

Examples of normalised shape vectors for the beer bottle can be seen in Fig. 4, for the different backgrounds. We note that the shape vectors are very consistent, with the exception of the poster table background in plot (g) where the centroid was incorrect, as the object was not properly detected and segregated.

The implicit encoding of local shape features is exemplified in Fig. 4(a) and can be seen if we consider the derivatives of the curves: positive or negative slopes represent bars/edges or curves, a zero slope being a perfectly circular curvature. Spikes and signal changes are zero-crossings of the second derivative and represent corners. In case of the beer bottle (Fig. 4a), one can recognise, following the contour anti-clockwise, the bottle's left side (0°), bottom (90°), right side (180°) and cap (270°), finally ending at the left side (360°). Hence, this shape coding scheme generalises our earlier gist model Martins et al. (2012) without using explicit geometric relations for encoding low-level features. A matrix of correlation plots between all vector pairs is shown in Fig. 5. The poster table vector was the only one that did not have a statistically significant positive correlation ($p < 0.05$) with the other vectors.

4. Object shape categorisation framework

For the experimental object categorisation setup, we used artificial neural networks to classify all 300 objects from the RGB-D Object Dataset (Lai et al., 2011), into 51 categories. Before detailing this process, we first introduce the classification rules used.

4.1. Classification rules

Generally, an object classification system can operate in either verification mode or in identification mode. In verification, the goal is to accept or reject an identity claim that is being presented to the system. In identification there is no identity claim; the system must find the object class which best matches the input object. In both cases there will be true and false positives and negatives, and the goal is to minimise the false ones. We will detail classification performance for both cases.

Verification mode. As mentioned above, this mode serves to verify the identity of an object whose representation is being presented to the system. The output feature vector $\boldsymbol{v}$ is a (trained) function of the actual input $\boldsymbol{u}$. For a trained neural network, $\boldsymbol{v}$ is the output of the network with dimension N, the number of known object classes, i.e., $N \triangleq 51$.

Feature vector $\boldsymbol{v}$ must be compared with a class template feature vector $\boldsymbol{v}_c$ if the claimed identity is U_c, with $c \in \{1, 2, \ldots N\}$. The class template feature vector $\boldsymbol{v}_c$ is the mean of a network's output responses from the training data of each class, for the different input data types of $\boldsymbol{u}_c$. The result R of the comparison is binary: either accept (1) or reject (0) the identity claim, considering a chosen acceptance threshold Δ (Štruc and Pavešić, 2010). This is formulated as

$$R \triangleq \begin{cases} 1, & \text{if } \xi(\boldsymbol{v}, \boldsymbol{v}_c) \geq \Delta \\ 0, & \text{otherwise,} \end{cases}$$

where $\xi(\cdot, \cdot)$ is a similarity function, for which we use the cosine similarity measure (Štruc and Pavešić, 2010),

$$\xi_{\cos}(\boldsymbol{v}, \boldsymbol{v}_c) \triangleq -\frac{\boldsymbol{v}^T \boldsymbol{v}_c}{\|\boldsymbol{v}\| \|\boldsymbol{v}_c\|}.$$

For experimental evaluation, the acceptance threshold Δ will be varied in order to determine false-rejection *vs*. false-acceptance curves.

Identification mode. A system running in this mode tries to identify an object by finding in the database the class template feature vector $\boldsymbol{v}_c$ that best matches the input feature vector $\boldsymbol{v}$, above an acceptance threshold Δ (Štruc and Pavešić, 2010):

$$U \triangleq \begin{cases} U_c, & \text{if } \; c = \underset{\kappa \in \{1,\ldots,N\}}{arg \max}[\xi_{\cos}(\boldsymbol{v}, \boldsymbol{v}_\kappa)] : \xi_{\cos}(\boldsymbol{v}, \boldsymbol{v}_\kappa) \geq \Delta \\ U_{N+1}, & \text{otherwise.} \end{cases}$$

Now U_{N+1} is the case when the input vector $\boldsymbol{v}$ cannot be matched to any of the $N \triangleq 51$ objects in the database. Since only cumulative

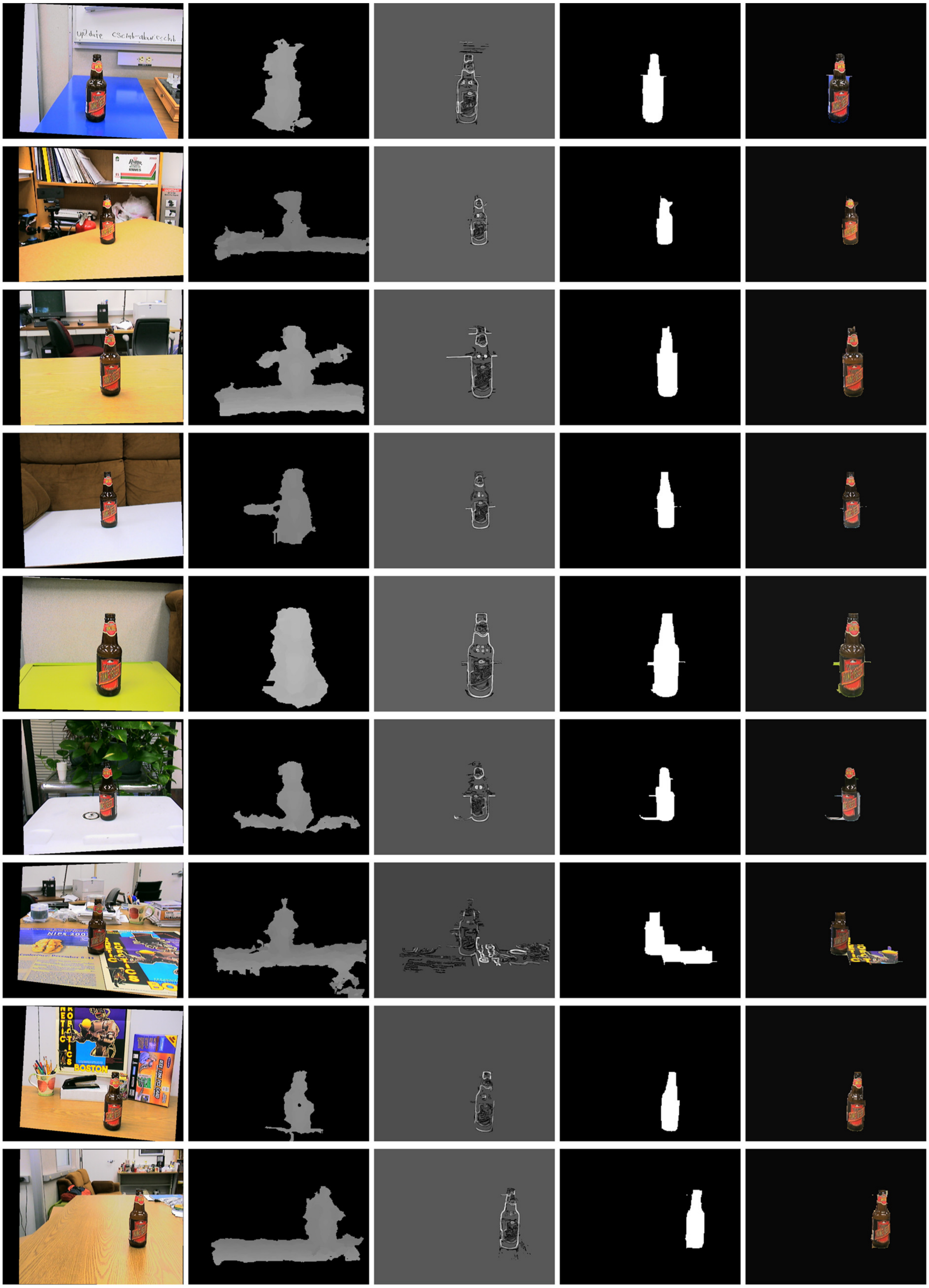
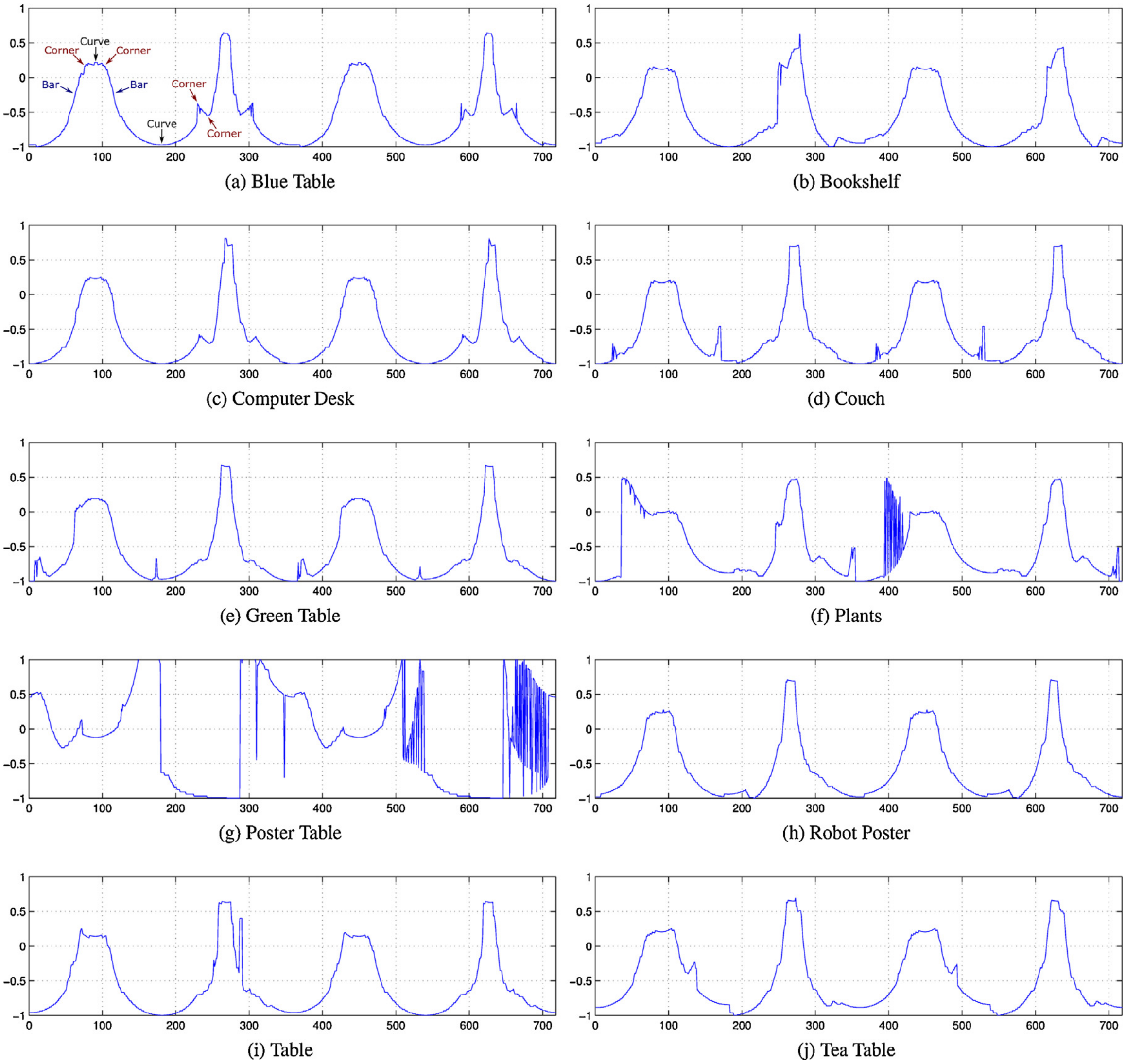


Fig. 3. Beer bottle shape extraction with the other nine different backgrounds of CSCLAB ID. *Left to right:* left image of the stereo pair; foreground object region disparity; object conspicuity; shape mask; and object segregation.



(a) Blue Table

(b) Bookshelf

(c) Computer Desk

(d) Couch

(e) Green Table

(f) Plants

(g) Poster Table

(h) Robot Poster

(i) Table

(j) Tea Table

Fig. 4. Beer bottle's 720-element normalised shape vectors $\boldsymbol{s}^N$, for the 10 different backgrounds. The first 360 elements of each vector correspond to the outer shape distances $\rho^{\max}$ and the next 360 elements to the inner shape distances $\rho^{\min}$.

match characteristic (CMC) curves will be measured in the experimental evaluation (see below), the acceptance threshold will be zero and therefore U_{N+1} will not be considered.

4.2. Experimental setup

For the experimental object categorisation setup, we used the RGB-D Object Dataset (Lai et al., 2011). The predominant reasons for using this dataset are: (a) its sheer volume of object data – 300 objects with full revolution data, taken at angles of approximately 30°, 45° and 60° above the horizon – a total of 250,000 RGB plus Depth frames, and (b) the availability of several published results by different authors, which allows for a quantitative comparison.

The RGB-D Object Dataset has 300 objects in 51 categories. We sampled the turntable data exactly as in Lai et al. (2011), using only every 5th video frame, for a total of 41,942 RGB-D images.[5] For every one of the 51 categories, we left the first object out for testing and used all the remaining ones for training. Thus, our training set, per experimental condition, consists of 34,921 images, while the test set counts 7021 images.

We applied seven experimental conditions on the training and test sets, each condition using a specific data type (i.e., luminance, colour, conspicuity, range, or shape), and combinations of these. Thus, we measured categorisation performance using: (1) luminance of the cropped objects; (2) colour; (3) conspicuity; (4) range; (5) shape; (6) shape plus range; and (7) shape plus luminance, conspicuity and range. For extracting the shape vectors we used the methodology and empirical parameters as detailed in the previous section.

4.3. Data pre-processing

Prior to classification, there are three pre-processing steps done for all training and test set images: (a) As RGB-D images are already cropped to the object area, there is no need for a detection step; they are just rescaled to 60 × 60 pixels (because each object crop has a different size).[6] (b) Luminance images I_L are created from the

[5] From here onwards we will use the term *range* instead of *disparity* to avoid confusion with the previous DEM-obtained disparity data, as it is now extracted directly from the dataset.

[6] This is often done in computer vision to reduce the number of neurons in a neural network classifier. It also has a biological background, as object templates stored in memory are normalised in size (DiCarlo et al., 2012).

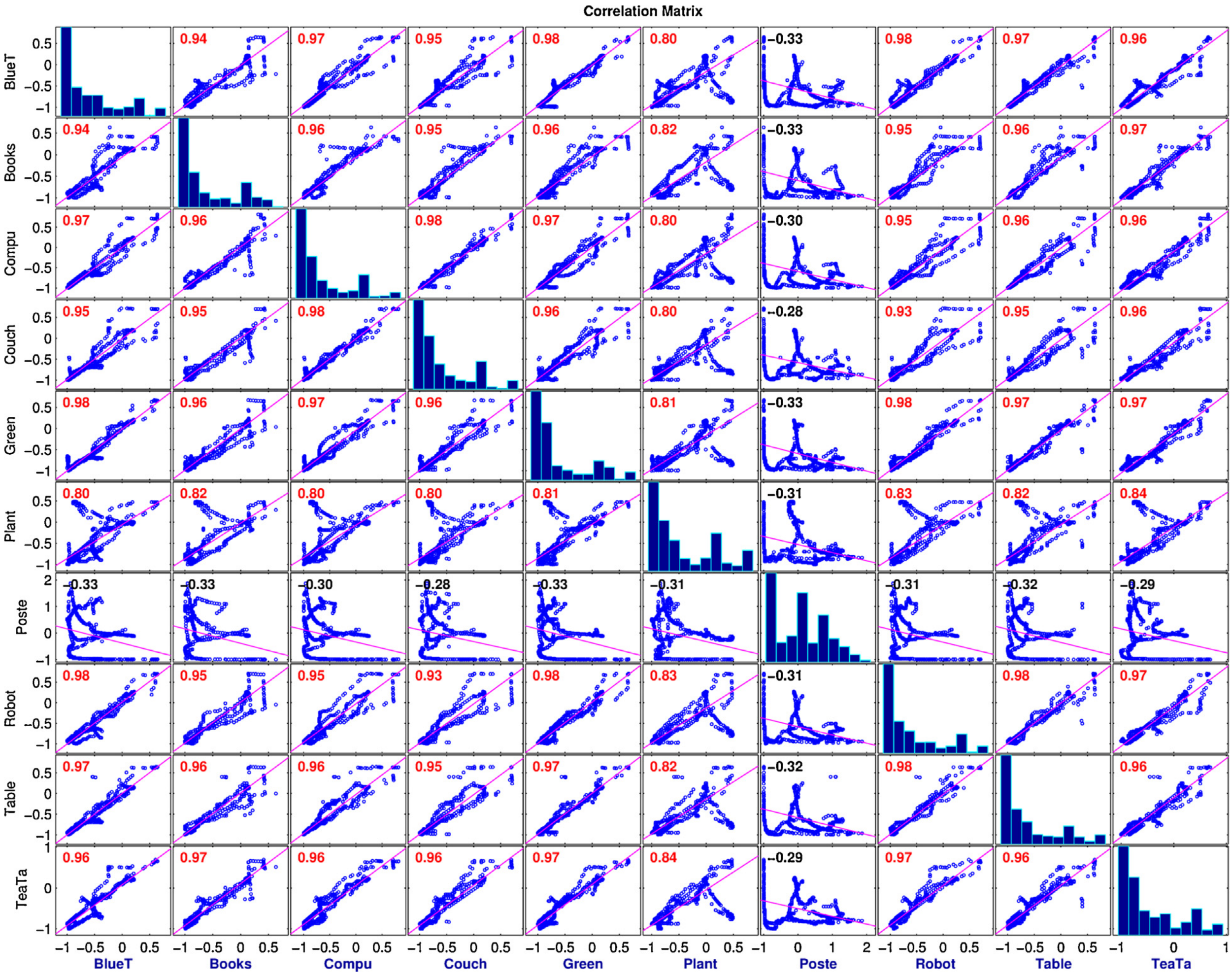


Fig. 5. Matrix of plots showing Pearson correlations between all pairs of the beer bottle's shape vectors. The main diagonal shows each vector's histogram, while the remaining cells show scatter plots of the vector pairs. Each cell has a least-squares reference line with slope equal to the Pearson's correlation coefficient. Significant positive correlations ($p < 0.05$) are highlighted in red. (For interpretation of the references to color in this legend, the reader is referred to the web version of the article.)

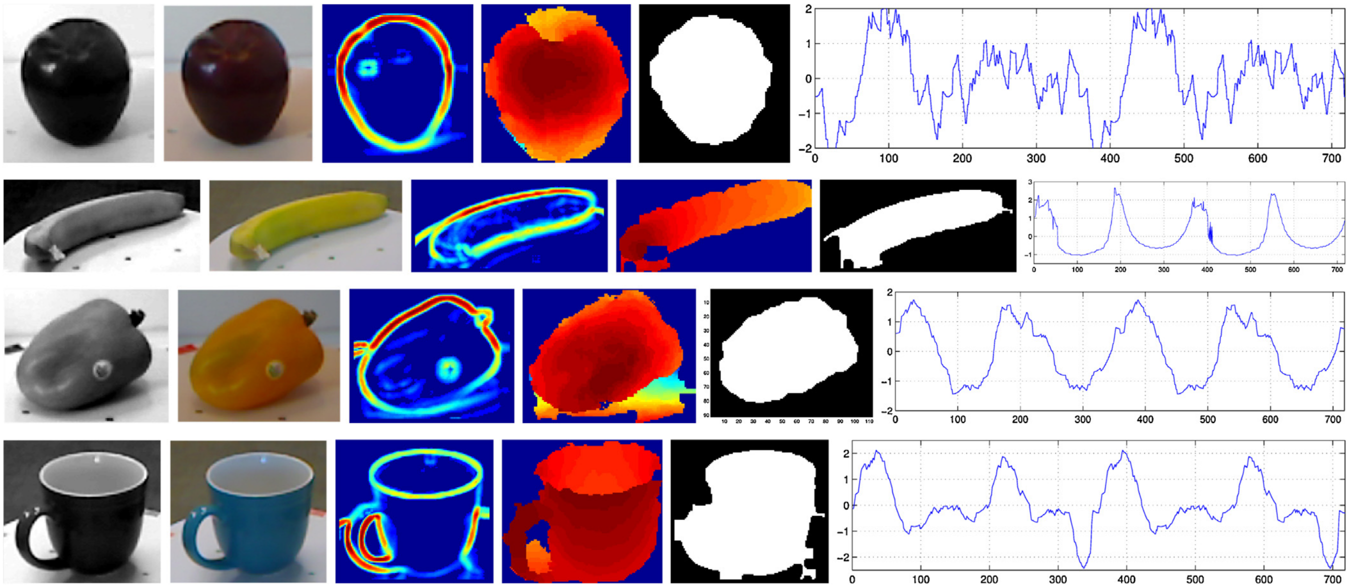


Fig. 6. Examples of RGB-D OD objects used for classification trials, with respective data types (luminance and colour images, conspicuity and range maps, binary masks and shape vectors extracted from each mask). *Top to bottom:* apple, banana, bellpepper and cup. (For interpretation of the references to color in this legend, the reader is referred to the web version of the article.)

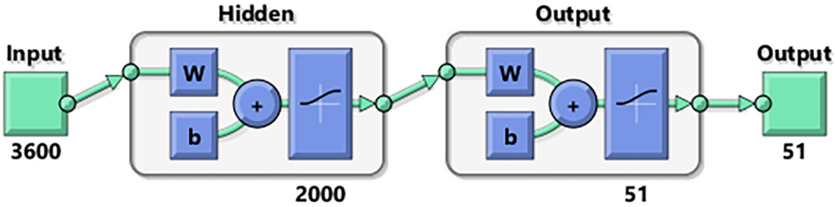


Fig. 7. Neural network topology used for RGB-D OD classification. Two-layer, feed-forward neural network with 2000-neuron log-sigmoid hidden layer and 51-neuron log-sigmoid output layer. The input vector is either of size 720 for shape vectors, 3600 (60 × 60 px) for image vectors, or the sum of the sizes for multi-modal classification.

original RGB images I_C, by reducing these to grayscale format. (c) All images are normalised by subtracting their mean and dividing by their standard deviation, which yields: luminance I_L^N, colour I_C^N, conspicuity C^N and range R^N maps. Shape vectors $\boldsymbol{s}^N$ are already normalised, so they are used with all their 720 components. Examples of each data type are shown in Fig. 6.

4.4. Classifiers

We applied artificial neural networks (NNs) as classifiers, which have a biological background. By using NNs we can: (a) prove that the framework can be completely implemented by applying biological principles, still obtaining good performance and (b) show that shape information can provide robust features for object categorisation.

A total of seven networks (one per experimental condition) were trained to classify the 51 object types of the training set. The general network topology is shown in Fig. 7. All networks are two-layer, feed-forward, using log-sigmoid hidden and output neurons, with the hidden layer composed of 2000 neurons (chosen after several empirical trials) and the output layer of 51 neurons. Training was done iteratively by resilient backpropagation, with as stopping criterion a maximum of 1000 epochs. The performance criterion was the mean-squared normalised error with a regularisation factor $\gamma = 0.2$ to avoid over-fitting.[7] This means that the NNs converged gradually towards the final solution, with the outputs of the networks not being binary: the output of a given object category is maximised and those of all other categories are smaller but not necessarily zero.

For each experimental condition, the input vector $\boldsymbol{u}$ of the respective network maps one neuron for each pixel position in I_L^N, I_C^N, C^N or R^N, for a total of 3600 input neurons (60 × 60 pixels); $\boldsymbol{s}^N$ maps to 720 neurons, ($\boldsymbol{s}^N$, R^N) to 4320 neurons and ($\boldsymbol{s}^N$, R^N, C^N, I_L^N) to 11,520 neurons.

5. Experimental results

5.1. Performance measures

For presenting performances of the different experimental setups, we measured standard error and recognition rates[8] which are commonly used in object categorisation and recognition research (Štruc and Pavešić, 2009, 2010).

Cumulative match characteristic (CMC). For class identification experiments we will show results in the form of recognition rates. We first computed the rank-one recognition (ROR) rate on the test set:

$$\text{ROR} \triangleq \frac{n_{ca}}{n_{ni}} 100\%,$$

where n_{ca} is the number of images assigned to the correct objects and n_{ni} is the total number of test images. ROR rates were complemented by the ranking beyond the first position, i.e., from rank 1 to rank τ. This yields a CMC curve that plots recognition rate by rank. In calculating the recognition rate for the τth rank, identification is considered successful if the correct identity is among the top τ results. CMC results are particularly useful for a local gist system, where the top τ matches can be used to bias scene categorisation. After fast gist vision, the top matches can then be scrutinised sequentially to increase certainty.

Detection error trade-off (DET). For class verification experiments we measured the false acceptance rate (FAR) and the false rejection rate (FRR), as well half total error rate (HTER). FAR and FRR are defined by

$$\text{FAR} \triangleq \frac{n_{ar}}{n_r} 100\%; \quad \text{FRR} \triangleq \frac{n_{ra}}{n_a} 100\%,$$

with n_{ar} the number of accepted illegitimate identity claims, n_r the number of all illegitimate identity claims, n_{ra} the number of rejected legitimate identity claims, and n_a the number of all legitimate identity claims. HTER is the average

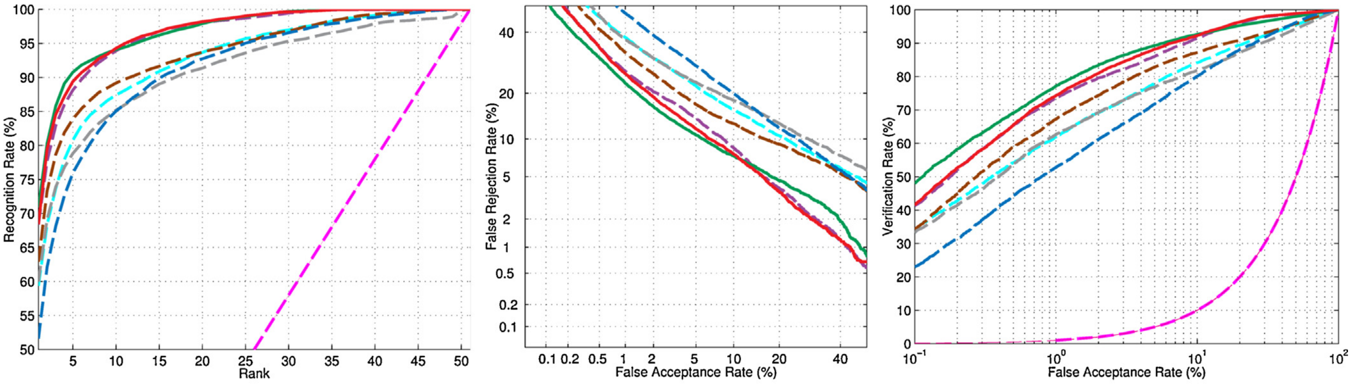
$$\text{HTER} \triangleq (\text{FAR} + \text{FRR})/2.$$

Both FAR and FRR depend on the value of the acceptance threshold Δ. When the one decreases, the other increases. To show the effect of Δ on FAR and FRR, the two error rates must be plotted against each other, for all possible values of the acceptance threshold, in the form of detection error trade-off (DET) curves. These relate FAR and FRR for different values of Δ on a scale defined by the inverse of a cumulative Gaussian density function (Štruc and Pavešić, 2010). A DET curve can be summarised by the equal error rate (EER), the point where FAR = FRR, with a lower value representing a more accurate result. Instead of the normal HTER, we will list HTER_{min}, which is the minimum possible HTER value, corresponding to the point of the DET curve closest to the origin, with $\text{HTER}_{\text{min}} \leq \text{EER}$. We will also include verification results at two FAR rates: from a moderate $\text{FAR}_{1\%}$ to a more stringent $\text{FAR}_{0.1\%}$.

Receiver operating characteristic (ROC). For class verification experiments, we also present results in terms of ROC curves that show the true acceptance rate, also known as verification rate or sensitivity (%), for a range of increasing FAR values (meaning decreasing specificity). Random-guess results will be plotted as dashed magenta lines. ROC curves are often summarised by the area under the curve: AUC. A larger AUC implies a better result. For example, even with a stringent false acceptance rate, the verification rate (the number of true positives) should still be high.

[7] The mean squared error will take into account the mean squared weights of the network, by $MSE_{REG} = \gamma MSE + (1 - \gamma) MSW$.

[8] Typical performance metrics generally use the term *recognition* independently of the actual task being categorisation or recognition. To avoid confusion we would like to note that categorisation is in fact recognising an object class or category.



(a) CMC (b) DET (c) ROC

Fig. 8. CMC, DET and ROC performance curves for the 51 object classes of RGB-D OD. Single modality results are given by dashed lines and joint results by solid lines. Shape in blue, range in purple, conspicuity in brown, luminance in grey, colour in cyan, shape and range in red, and shape, range, conspicuity and luminance in green. CMC and ROC random-guess rates are dashed in magenta. (For interpretation of the references to color in this legend, the reader is referred to the web version of the article.)

Table 1
NN RGB-D OD performance results (51 object classes).

Data type	Identification	Verification				
	ROR$_{\%}$	EER$_{\%}$	AUC$_{\%}$	HTER$_{\text{min}\%}$	FAR$_{1\%}$	FAR$_{0.1\%}$
Shape	51.6	15.1	92.4	14.8	52.7	23.3
Range	68.4	9.2	96.7	9.0	73.5	41.2
Conspicuity	62.9	11.8	93.7	10.8	67.3	34.0
Luminance	60.0	15.1	91.6	13.8	62.7	33.2
Colour	59.4	13.6	92.7	12.8	61.9	34.1
Shape + range	68.5	8.6	**97.0**	8.2	74.3	41.5
Shape + range + conspicuity + luminance	**70.9**	**8.1**	96.2	**7.8**	**77.1**	**48.3**

Note: Best results in each column are printed in bold.

5.2. Performance assessments

Performance curves and data for the experimental setups are shown in Fig. 8 and in Table 1. Different curves are used for the seven experimental conditions: using only $\boldsymbol{s}^N$ vectors (shape, as blue dashed lines), R^N maps (range, as purple dashed lines), C^N maps (conspicuity, as brown dashed lines), I_L^N images (luminance, as grey dashed lines), I_C^N images (colour, as cyan dashed lines) and using a combination of either ($\boldsymbol{s}^N$, R^N) (shape and range, as red solid lines) or ($\boldsymbol{s}^N$, R^N, C^N, I_L^N) (shape, range, conspicuity and luminance, as green solid lines). CMC and ROC random-guess rates are represented by dashed magenta lines.

Identification of an object class (categorisation) based on shape alone resulted in a ROR rate of 51.6%, quickly rising to a rank-5 rate around 77%, which is promising for a bio-inspired proto-object categorisation system. Verification rates are lower than for the other experimental conditions, showing that shape alone is not discriminative enough for class verification purposes at very low error rates, but it still achieved around 80% at FAR$_{10\%}$ error. Overall, shape results are impressive considering that the size of the shape vector is 1/5th of the other feature spaces.

Range was the most discriminative modality, with the top single-modality results in both class identification and verification (see Table 1). Colour achieved slightly better CMC results than luminance on class identification (except for the ROR rate) and both achieved similar rates on verification. Conspicuity was overall able to achieve better results than both, with a bigger margin on verification trials. The best performance was obtained when combining the four modalities, although the shape plus range results are very close. When using four modalities, the rank-5 rate of 77% for shape increases to 91%.

We compare our ROR categorisation results with those of other authors in Table 2. Our Proto-NN results are shown (for each column) using: (I) only shape vectors, (II) only range maps and (III)

Table 2
RGB-D OD categorisation ROR$_{\%}$ rate comparison.

Classifier	Employed features for shape/vision	Shape	Vision	All
Proto-NN	Shape vectors only | range maps only | all data types	51.6	68.4	70.9
Linear SVM (Lai et al., 2011)	Spin images + efficient match Kernels (EMK) + random Fourier sets + width + depth + height | SIFT + texton histogram + colour histogram	53.1	74.3	81.9
Random forest (Lai et al., 2011)	(Same)	66.8	74.7	79.6
kSVM (Lai et al., 2011)	(Same)	64.7	74.5	83.8
SVM (Bo et al., 2011)	3D shape + physical size of the object + depth edges + gradients + Kernel PCA + local binary patterns + multiple depth kernels	78.8	77.7	86.2
CKM (Blum et al., 2012)	SURF interest points	–	–	86.4
CNN-RNN (Socher et al., 2012)	ZCA whitening + softmax classifier	78.9	80.8	86.8
SP + HMP (Bo et al., 2013)	Surface normals	81.2	82.4	87.5

Note: Results ordered by average performance considering all three columns.

four data modalities. Also shown are the ROR results of Lai et al. (2011), Bo et al. (2011), Blum et al. (2012), Socher et al. (2012) and Bo et al. (2013).

The shape column shows that the proposed method has almost reached the performance of a linear SVM in Lai et al. (2011), with a very small −1.5% difference. These authors used a state-of-the-art computer vision algorithm for shape categorisation, based on features extracted from the 3D location of each depth pixel, relying on spin-images that capture the spatial distribution of a randomly sampled set of 3D points, expressing them into 16 × 16 histograms. These are used to compute efficient match kernel (EMK) features using random Fourier sets and principal component analysis (PCA) to arrive at a 2703-dimensional shape descriptor (3.8 times larger than our 720).

In the vision column (range map only), our rank-one result based on 3600 features was below the worst computer vision algorithm, which is again a linear SVM in Lai et al. (2011), at a −5.9% difference. For categorisation the authors used SIFT descriptors on a dense grid of 8 × 8 cells with EMK features at two scales, followed by PCA, achieving a 1500-dimensional vector to be used along with texton histograms (from oriented Gaussian filter responses), a colour histogram and the mean/standard deviation of each colour channel, as visual features (the total number is not specified by the authors).

The all column shows that our rank-one result did not improve much when combining four data modalities. We only reached 70.9%, a −8.7% difference to the random Forest classifier in Lai et al. (2011).

It is unfortunate that all authors only published ROR data, so we cannot compare the evolution in classifier performance (neither CMC nor DET/ROC rates), which would allow for a more detailed view of the complete system performance. For example, we can see in Fig. 8(a) that recognition rates quickly increase after rank one, with shape results achieving around 76% at rank-5 and 85% at rank-10. Even the 70.9% ROR for four modalities is able to reach around 91% at rank-5. These are very important considerations for a local gist system, where global scene categorisation can be primed from several close object categories, without requiring precise rank-one results (e.g., the common object families in an office will be very different from those in a corridor).

6. Discussion

In general, the performance results clearly emphasise the role of shape and 3D information in object categorisation, more than the obvious benefit of both being invariant to lighting conditions. Both yield a strong structural representation suitable to classify objects with good accuracy, quickly surpassing a rank-5 recognition rate of 76% using shape data and 89% for shape plus range data. This is perhaps more than enough for bootstrapping gist vision, since object categorisation can be done by the visual system in parallel streams – quickly classifying just a few familiar objects in each scene is often enough for hugely biasing scene recognition. We also note that conspicuity features were able to outperform both luminance and colour data, highlighting their discriminative capabilities. In further research it makes sense to expand the categorisation scheme, integrating additional low-level input features, such as lines/edges, textures and keypoints that are readily available from simple, complex and end-stopped cells in V1/V2 (Rodrigues and du Buf, 2006, 2009; Martins et al., 2012).

Our ROR shape-only result (51.6%) indicates that the proposed proto-object shape categorisation method is within reach of much more computationally advanced and complex methods, based on state-of-the-art spin images with EMK features (53.1%) (Lai et al., 2011). Perhaps even more important is that a system which employs cortical neuronal processes, i.e., which can be thought of as mimicking part of our visual system, can be applied to a real-world problem in computer vision.

The proposed method fits into a broader cortical architecture as is shown in Fig. 9, which highlights the possible neural pathways which link different processes. This architecture relies on both standard and non-standard retinal ganglion cells. The *non-standard* pathway (Fig. 9, bottom right) can serve to quickly bootstrap the low-level gist module, by using specific, hard-coded, shape descriptors (*corners*, *bars* and *curves*) to feed a *geometric shape classifier*, which can be constructed using higher-level grouping cells (Martins et al., 2012). Since low-level geometry information has already been extracted, it is therefore available for obtaining local object gist, e.g., providing cues which are used for a first and fast selection of possible object categories in memory (Bar et al., 2006). This is a purely bottom-up and data-parallel process for bootstrapping the serial object categorisation and recognition processes, which are controlled by top-down attention. Recent research also suggests that we actually categorise objects before we have segregated them, or that both processes occur in parallel. This means that by the time we realise that we are looking at something, our brain already knows what that thing is (Oliva and Torralba, 2006). Therefore, Rensink (2000) proposed a non-attentional "scene schema" consisting of concurrent spatial-layout and gist subsystems which both drive attentional object recognition, all employing "proto-objects" resulting from low-level vision. Similarly, Yanulevskaya et al. (2013) also focused on salient proto-object detection within an object-based attention theory. However, gist vision addressed so far mostly concerns global gist of entire scenes (Bar, 2004; Siagian and Itti, 2007; Ross and Oliva, 2010; Rodrigues and du Buf, 2011; Terzić et al., 2013). Global scene gist can be used to bias – select or exclude – object templates in memory in the matching process: when in a classroom it is not very likely that we see a horse. But global gist lacks localisation. On the other hand, when seeing a horse it is not very likely that we are in a classroom. Local object gist has the advantage of solving, or at least contributing to, the spatial-layout subsystem as proposed by Rensink (2000). Although both global and local gist can determine context, probably with a straight relation between them, local gist can solve important problems like a first and fast object categorisation, localisation and segregation, the latter being related to figure-ground organisation (Craft et al., 2007).

A similar low-level attentional view is described by Martin and von der Heydt (2013), who measured spike time correlations in monkey visual cortex. They concluded that specific grouping cells in V1/V2 were able to specifically enhance the activity of neurons whose receptive fields fit their grouping templates, linking neurons to "proto-object structures." In this context, we expect that global gist features will suffice for initial discrimination between very different scene types (with regard to spatial layout, like a "forest" *vs.* a "city") but will severely lack detail in more similar scenes (like an "office" *vs.* a "classroom"). Here local gist can be of great benefit – since most man-made objects tend to possess well-defined geometric shapes – when employed in conjunction with global scene properties (Rodrigues and du Buf, 2011). This view is also reinforced by Groen et al. (2013), who concluded that gist seems to depend on two stages: an early, automatic stage, where local-contrast responses present in the LGN or V1 seem to play a very important role, followed by a later, task-dependent stage.

Along the *standard* path (Fig. 9, left) binocular simple and complex cells are used to create a *disparity map* (Martins et al., 2015), while their monocular versions can be used for the *line/edge map* (Rodrigues et al., 2012), *conspicuity map* (Martins et al., 2012) and *keypoint map* (Rodrigues and du Buf, 2011). According to the attentional coherence theory by Rensink (2000), low-level proto-object shapes are continually formed, rapidly and in parallel across the visual field – they are volatile, lacking strong coherence until being

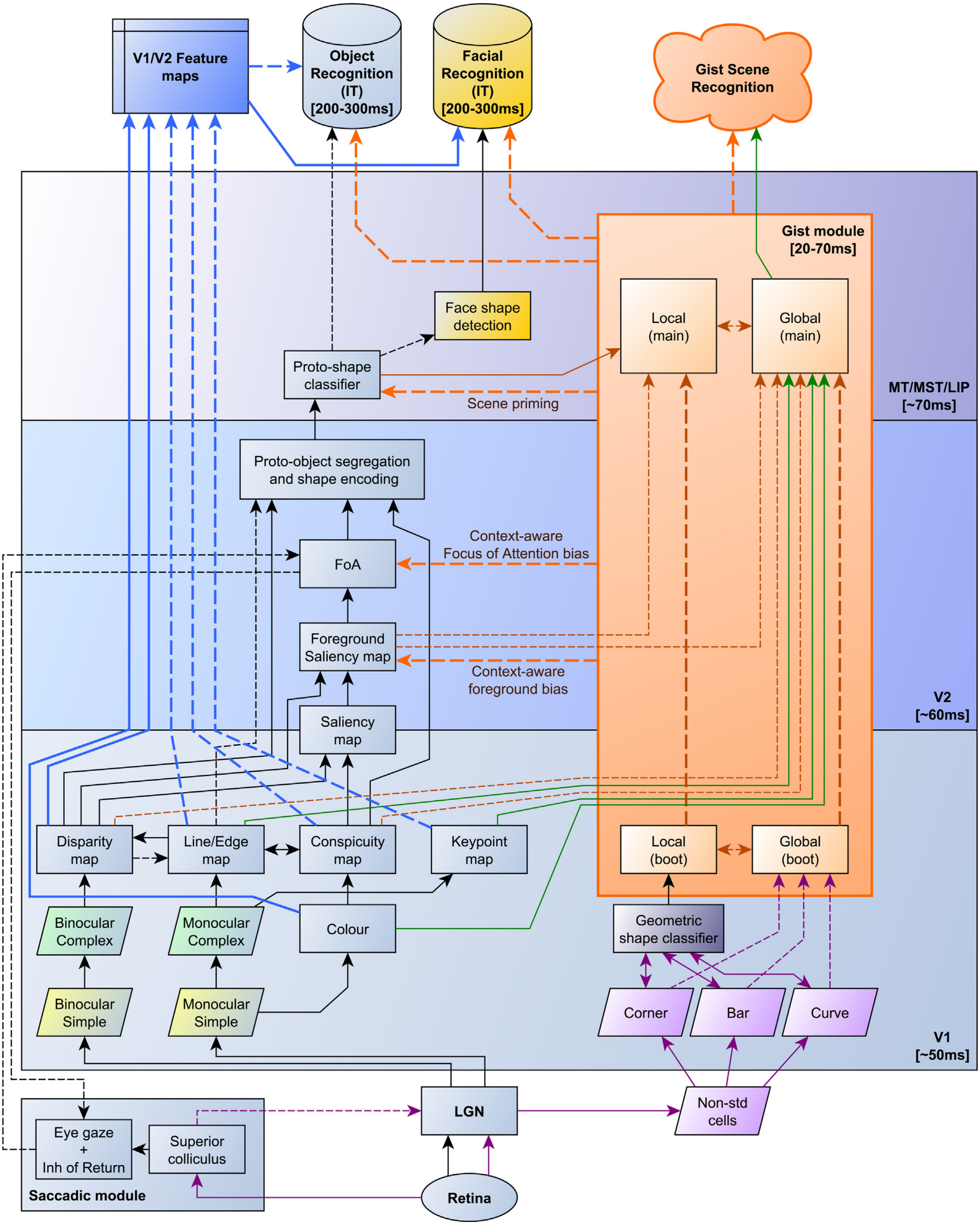


Fig. 9. Proposed cortical architecture for gist, object and face recognition. Solid arrows represent previous research and dashed arrows represent expected links for future research. Green arrows represent the low-level global gist architecture developed by Rodrigues and du Buf (2011).

stabilised by FoA-gaze, afterwards dissolving when FoA is released. In our model, we postulate that available disparity and conspicuity information, when combined, is able to quickly highlight all important objects and to resolve border ownership of the outlines of objects. Higher-level, oriented grouping cells can encode the distance from the centre of the object to its border, for very specific orientations. In our case, we implemented two populations of 360 of these cells, with orientations separated by one degree (a total of 720 cells). The reason for using two populations for a seemingly similar task is that borders are seldom unique: complex shapes can have several border transitions at certain orientations, i.e., they are not star-shaped polygons. To resolve this problem, one of the populations encodes the distance to the first, closest border near the centre of the object, whereas the second population encodes the farthest border. This allowed for significant resilience in categorising complex shapes. These population responses can serve as inputs to a simple, feed-forward *proto-shape classifier*, probably residing in area LIP in the dorsal pathway (Konen and Kastner, 2008; Janssen et al., 2008). LIP also shows shape activation times of 62 ms (Lehky and Sereno, 2007), well within global gist recognition times and it has access (via areas V5/MT and MST), to the superior colliculus for eye and head control (Gottlieb, 2007), crucial for FoA. This makes LIP a prime candidate area for integrating low-level attention with a proto-object categorisation role.

Salient foreground regions can also serve as inputs for FoA, which can use available proto-object shapes. The ventral stream (V4 and PIT) can further refine shape information and bootstrap either object recognition or face recognition, since we know that the cortex employs a dedicated pathway for face processing (Biederman and Kalocsai, 1997).

Gist is expected to have a significant biasing effect throughout this whole process, which is an important area of interest for further research. For example, scene context can influence the categorisation of objects or faces, by biasing the range of familiar matches related to each context. Also, when the brain establishes a background/foreground split in a scene, gist can bias the split based on the scene context (i.e., a forest *vs.* an office space) helping to choose, for each, the best close-to-far range for foreground depth. Later it can influence FoA by increasing the priority of salient shapes depending on context: a very close “bear” shape in a “forest” will get top priority!

Acknowledgements

This work was partly supported by the Portuguese Foundation for Science and Technology (FCT), LARSyS FCT [UID/EEA/50009/2013], EU project NeuroDynamics (FP7-ICT-2009-6, PN: 270247), FCT project SparseCoding (EXPL/EEI-SII/1982/2013) and FCT PhD grant to author J.A.M. (SFRH-BD-44941-2008). We thank the anonymous reviewers for their helpful comments.

References

Bar, M., Kassam, K.S., Ghuman, A.S., Boshyan, J., Schmid, A.M., Schmidt, A.M., Dale, A.M., Hämälä inen, M.S., Marinkovic, K., Schacter, D.L., Rosen, B.R., Halgren, E., 2006. Top-down facilitation of visual recognition. In: Proc. Natl. Acad. Sci. U. S. A., vol. 103, pp. 449–454, http://dx.doi.org/10.1073/pnas.0507062103 http://www.pnas.org/content/103/2/449

Bar, M., 2004. Visual objects in context. Nat. Rev. Neurosci. 5, 619–629.

Biederman, I., Kalocsai, P., 1997. Neurocomputational bases of object and face recognition. Philos. Trans. R. Soc. Biol. Sci. 352, 1203–1219.

Blum, M., Wulfing, J., Riedmiller, M., 2012. A learned feature descriptor for object recognition in RGB-D data. In: 2012 IEEE Int. Conf. Robot. Autom, pp. 1298–1303, http://dx.doi.org/10.1109/ICRA.2012.6225188 http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6225188

Bo, L., Ren, X., Fox, D., 2011. Depth kernel descriptors for object recognition. In: 2011 IEEE/RSJ Int. Conf. Intell. Robot. Syst, pp. 821–826, http://dx.doi.org/10.1109/IROS.2011.6095119 http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6095119

Bo, L., Ren, X., Fox, D., 2013. Unsupervised feature learning for RGB-D based object recognition. In: 13th Int. Symp. Exp. Robot., vol. 88, pp. 387–402, http://dx.doi.org/10.1007/978-3-319-00065-7, http://dblp.uni-trier.de/db/conf/iser/iser2012.html#BoRF12, http://link.springer.com/10.1007/978-3-319-00065-7

Craft, E., Schütze, H., Niebur, E., von der Heydt, R., 2007. A neural model of figure-ground organization. J. Neurophysiol. 97, 4310–4326, http://dx.doi.org/10.1152/jn.00203.2007 http://www.ncbi.nlm.nih.gov/pubmed/17442769

Davies, E.R., 2004. Machine Vision: Theory, Algorithms, Practicalities. Elsevier http://books.google.com/books?hl=en&lr=&id=uY-Z3vORugwC&pgis=1

DiCarlo, J.J., Zoccolan, D., Rust, N.C., 2012. How does the brain solve visual object recognition? Neuron 73, 415–434, http://dx.doi.org/10.1016/j.neuron.2012.01.010 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3306444&tool=pmcentrez&rendertype=abstract

Farivar, R., 2009. Dorsal-ventral integration in object recognition. Brain Res. Rev. 61, 144–153 http://www.ncbi.nlm.nih.gov/pubmed/19481571

Gopalakrishna, M.T., Ravishankar, M., Rameshbabu, D.R., 2014. Multiple moving object recognitions in video based on log Gabor-PCA approach. In: Proc. Second Int. Symp. Intell. Informatics. Recent Advances in Intelligent Informatics, vol. 235. Springer, Mysore, India, pp. 93–100, http://link.springer.com/chapter/10.1007/978-3-319-01778-5_10.

Gottlieb, J., 2007. From thought to action: the parietal cortex as a bridge between perception, action, and cognition. Neuron 53, 9–16, http://dx.doi.org/10.1016/j.neuron.2006.12.009 http://www.ncbi.nlm.nih.gov/pubmed/17196526

Groen, I., Ghebreab, S., Lamme, V., Scholte, H.S., 2013. Two stages in scene gist processing revealed by evaluating summary statistics with single-image ERPs. J. Vis. 13, 1059, http://dx.doi.org/10.1167/13.9.1059, http://www.journalofvision.org/content/13/9/1059.abstract, http://www.journalofvision.org/lookup/doi/10.1167/13.9.1059

Janssen, P., Srivastava, S., Ombelet, S., Orban, G.A., 2008. Coding of shape and position in macaque lateral intraparietal area. J. Neurosci. 28, 6679–6690, http://dx.doi.org/10.1523/JNEUROSCI.0499-08.2008 http://www.ncbi.nlm.nih.gov/pubmed/18579742

Johnson, A.E., Hebert, M., 1998. Efficient multiple model recognition in cluttered 3-D scenes. In: Proc. 1998 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog, pp. 671–677 http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=698676

Kogo, N., Wagemans, J., 2013. The emergent property of border-ownership and the perception of illusory surfaces in a dynamic hierarchical system. Cogn. Neurosci. 4, 54–61, http://dx.doi.org/10.1080/17588928.2012.754750 http://www.ncbi.nlm.nih.gov/pubmed/24073704

Konen, C.S., Kastner, S., 2008. Two hierarchically organized neural systems for object information in human visual cortex. Nat. Neurosci. 11, 224–231 http://www.ncbi.nlm.nih.gov/pubmed/18193041

Lai, K., Bo, L., Ren, X., Fox, D., 2011. A large-scale hierarchical multi-view RGB-D Object Dataset. In: Proc. IEEE Int. Conf. Robot. Autom, pp. 1817–1824, http://dblp.uni-trier.de/db/conf/icra/icra2011.html#LaiBRF11, http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5980382.

Lehky, S.R., Sereno, A.B., 2007. Comparison of shape encoding in primate dorsal and ventral visual pathways. J. Neurophysiol. 97, 307–319, doi:10.1152/jn.00168.2006.

Lu, Y., Rasmussen, C., 2012. Simplified Markov random fields for efficient semantic labeling of 3D point clouds. In: IEEE/RSJ Int. Conf. Intell. Robot. Syst, pp. 2690–2697, http://dx.doi.org/10.1109/IROS.2012.6386039 http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6386039

Martin, A., von der Heydt, R., 2013. Firing synchrony between neurons reveals proto-object representation in monkey visual cortex. J. Vis. 13, 289, http://dx.doi.org/10.1167/13.9.289

Martins, J.A., Rodrigues, J.M.F., du Buf, J.M.H., 2009. Focus of attention and region segregation by low-level geometry. In: Proc. Fourth Int. Conf. Comput. Vis. Theory Appl. In: Ranchordas, A., Araújo, H. (Eds.), INSTICC Press, Lisbon, Portugal, pp. 267–272, doi:http://hdl.handle.net/10400.1/879.

Martins, J.A., Rodrigues, J.M.F., du Buf, J.M.H., 2011. Disparity energy model using a trained neuronal population. In: Proc. IEEE Int. Symp. Signal Proc. Inf. Technol., Bilbao, Spain, pp. 287–292, http://dx.doi.org/10.1109/ISSPIT.2011.6151575, http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6151575, http://hdl.handle.net/10400.1/2078

Martins, J.A., Farrajota, M., Lam, R., Rodrigues, J.M.F., Terzic, K., du Buf, J.M.H., 2012. A disparity energy model improved by line edge and keypoint correspondences. In: Proc. 35th Eur. Conf. Vis. Percept., vol. 41, suppl., Alghero, Italy, p. 76 http://w3.ualg.pt/jrodrig/papers_pdf/ecvp2012.pdf

Martins, J.A., Rodrigues, J.M.F., du Buf, J.M.H., 2012. Local object gist: meaningful shapes and spatial layout at a very early stage of visual processing. Gestalt Theory 34, 349–380, doi:http://hdl.handle.net/10400.1/2170, http://gth.krammerbuch.at/content/vol34-issueheft3-4

Martins, J.A., Rodrigues, J.M.F., du Buf, J.M.H., 2015. Luminance, colour, viewpoint and border enhanced disparity energy model. PLoS ONE 10, e0129908, http://dx.doi.org/10.1371/journal.pone.0129908

Murphy-Chutorian, E., Triesch, J., 2005. Shared features for scalable appearance-based object recognition. In: Proc. 7th IEEE Work. Appl. Comput. Vis., vol. 1, pp. 16–21, doi:10.1109/ACVMOT.2005.109, http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4129454.

Oliva, A., Torralba, A., 2006. Building the gist of a scene: the role of global image features in recognition. Prog. Brain Res. Vis. Percept. 155, 23–26.

Quigley, M., Batra, S., Gould, S., Klingbeil, E., Le, Q., Wellman, A., Ng, A.Y., 2009. High-accuracy 3D sensing for mobile manipulation: improving object detection and door opening. In: Proc. IEEE Int. Conf. Robot. Autom, pp. 2816–2822, doi:10.1109/ROBOT.2009.5152750, http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=5152750

Raudies, F., Ringbauer, S., Neumann, H., 2013. A bio-inspired, computational model suggests velocity gradients of optic flow locally encode ordinal depth at surface borders and globally they encode self-motion. Neural Comput. 25, 2421–2449, http://dx.doi.org/10.1162/NECO_a_00479 http://www.ncbi.nlm.nih.gov/pubmed/23663150

Rensink, R., 2000. The dynamic representation of scenes. Vis. Cogn. 7, 17–42.

Rodrigues, J., du Buf, J., 2006. Multi-scale keypoints in V1 and beyond: object segregation scale selection saliency maps and face detection. BioSystems 2, 75–90 http://sapientia.ualg.pt/handle/10400.1/181

Rodrigues, J.M.F., du Buf, J.M.H., 2009. Multi-scale lines and edges in V1 and beyond: brightness object categorization and recognition and consciousness. BioSystems

95, 206–226, http://dx.doi.org/10.1016/j.biosystems.2008.10.006 http://www.sciencedirect.com/science/article/pii/S0303264708002372

Rodrigues, J.M.F., du Buf, J.M.H., 2011. A cortical framework for scene categorization. In: Proc. Int. Conf. Comput. Vis. Theory Appl. (VISAPP 2011), Vilamoura, Portugal, pp. 364–371 https://sapientia.ualg.pt/handle/10400.1/889

Rodrigues, J.M.F., Martins, J.A., Lam, R., du Buf, J.M.H.,2012. Cortical multiscale line-edge disparity model. In: Proc. Int. Conf. Image Anal. Recognit. Springer LNCS 7324, Aveiro, Portugal, pp. 296–303 http://w3.ualg.pt/jrodrig/papers_pdf/iciar2012.pdf

Ross, M.G., Oliva, A., 2010. Estimating perception of scene layout properties from global image features. J. Vis. 10, 1–25 http://journalofvision.org/10/1/2/

Siagian, C., Itti, L., 2007. Rapid biologically-inspired scene classification using features shared with visual attention. IEEE Tr. Robot. 29, 300–312.

Socher, R., Huval, B., Bath, B., Manning, C.D., Ng, A.Y., 2012. Convolutional-recursive deep learning for 3D object classification. In: Adv. Neural Inf. Process. Syst, pp. 665–673 http://machinelearning.wustl.edu/mlpapers/papers/NIPS2012_0304

Štruc, V., Pavešić, N., 2009. Gabor-based kernel partial-least-squares discrimination features for face recognition. Informatica 20, 115–138 https://dl.acm.org/citation.cfm?id=1516709

Štruc, V., Pavešić, N., 2010. The complete Gabor–Fisher classifier for robust face recognition. EURASIP J. Adv. Signal Process 2010, 847680, http://dx.doi.org/10.1155/2010/847680 http://asp.eurasipjournals.com/content/2010/1/847680

Terzić, K., Lobato, D., Saleiro, M., Martins, J.A., Farrajota, F., Rodrigues, J.M.F., du Buf, J.M.H., 2013. Biological models for active vision: towards a unified architecture. In: Comput. Vis. Syst. – Spec. Issue 12. Springer LNCS, vol. 7963., pp. 113–122, http://dx.doi.org/10.1007/978-3-642-39402-7_12 http://hdl.handle.net/10400.1/3395

Yanulevskaya, V., Uijlings, J., Geusebroek, J.-M., 2013. Salient object detection: from pixels to segments. Image Vis. Comput. 31, 31–42, http://dx.doi.org/10.1016/j.imavis.2012.09.009 http://linkinghub.elsevier.com/retrieve/pii/S0262885612001795