Shape-Based Measures Improve Scene Categorization

Morteza Rezanejad, John Wilder, Dirk B. Walther, Allan D. Jepson, Sven Dickinson, Kaleem Siddiqi

Abstract—Converging evidence indicates that deep neural network models that are trained on large datasets are biased toward color and texture information. Humans, on the other hand, can easily recognize objects and scenes from images as well as from bounding contours. Mid-level vision is characterized by the recombination and organization of simple primary features into more complex ones by a set of so-called Gestalt grouping rules. While described qualitatively in the human literature, a computational implementation of these perceptual grouping rules is so far missing. In this article, we contribute a novel set of algorithms for the detection of contour-based cues in complex scenes. We use the medial axis transform (MAT) to locally score contours according to these grouping rules. We demonstrate the benefit of these cues for scene categorization in two ways: (i) Both human observers and CNN models categorize scenes most accurately when perceptual grouping information is emphasized. (ii) Weighting the contours with these measures boosts performance of a CNN model significantly compared to the use of unweighted contours. Our work suggests that, even though these measures are computed directly from contours in the image, current CNN models do not appear to extract or utilize these grouping cues.

Index Terms—Scene Categorization, Shape Based Measures, Gestalt Grouping Cues, Scene Perception, Contour Geometry, Medial Axis Transform.

1 Introduction

Present convolutional neural network (CNN)-based computer vision systems offer competitive recognition performance for various tasks. Although the achievements of CNN-based algorithms have changed the landscape of computer vision and machine learning in recent years, these models still lack some of the key capabilities that the human visual system has. Current deep neural network models are typically extremely data-hungry and do not necessarily represent all structural visual cues; rather, they extract and match appearance-based features to optimize their performance. This contrasts with human visual perceptual capabilities. As an example, while a computer vision system may need to train on hundreds of images of a cat, a child can learn this abstract category from a few example sketches of their outlines [1]. Such observations highlight a role for the exploitation of other visual cues such as shape-based ones in computer vision-based systems. These cues may lead us to better perceive the abstract form of a visual image and to efficiently and robustly generalize that abstract form to new exemplars of the same category. What mechanisms does the human visual system use in order to organize this potentially highly complex visual information to support high-level visual reasoning? And what can we learn from these principles for improving the performance of artificial vision systems?

While neural network models have provided a substantial boost in recognition and categorization performance, their complexity has prevented us from understanding how visual features are organized and how such organization is exploited in a meaningful way. On the other hand, the human visual system can readily exploit those visual features in both real-world and abstract scenarios, including the recognition or categorization of objects or entire scenes. In this article, we focus on the scene categorization problem and explore whether convolutional neural network architectures can benefit from organization principles that are inspired by biological vision systems (see Fig. 1).

Scene categorization cannot be easily disentangled from the recognition of objects, since scene classes are often defined by a collection of objects in context. A beach scene, for example, would typically contain umbrellas, beach chairs, and people in bathing suits, all of which are situated next to a body of water. A street scene might have roads with cars, cyclists, and pedestrians as well as buildings along their sides. How might computer vision systems tackle this problem of organizing objects and object parts to support scene categorization?

In human vision, perceptual organization is thought to be affected by a set of heuristic organizing rules originating from Gestalt psychology [3]. Such rules posit that visual elements ought to be grouped together if they are, for instance, similar in appearance, in close proximity, or if they are symmetric or parallel to each other. Originally developed as ad-hoc heuristics, these rules have been validated empirically, even though their precise neural mechanisms remain elusive.

Perceptual organization cues, such as those based on symmetry, are thought to aid in high-level visual tasks, such

- *M. Rezanejad, J. Wilder, D. B. Walther are with the Department of Psychology, University of Toronto, 100 St. George Street, 4^{th} Floor, Sidney Smith Hall, Toronto, ON M5S 3G3, Canada.*
- *A. D. Jepson, S. Dickinson are with the Department of Computer Science, University of Toronto, 40 St. George Street, Room 4283 Toronto, ON M5S 2E4, Canada.*
- *K. Siddiqi is with the School of Computer Science and the Centre for Intelligent Machines, McGill University, McConnell Engineering, 3480 University Street, Montreal, QC H3A 0E9, Canada E-mail: siddiqi@cim.mcgill.ca*

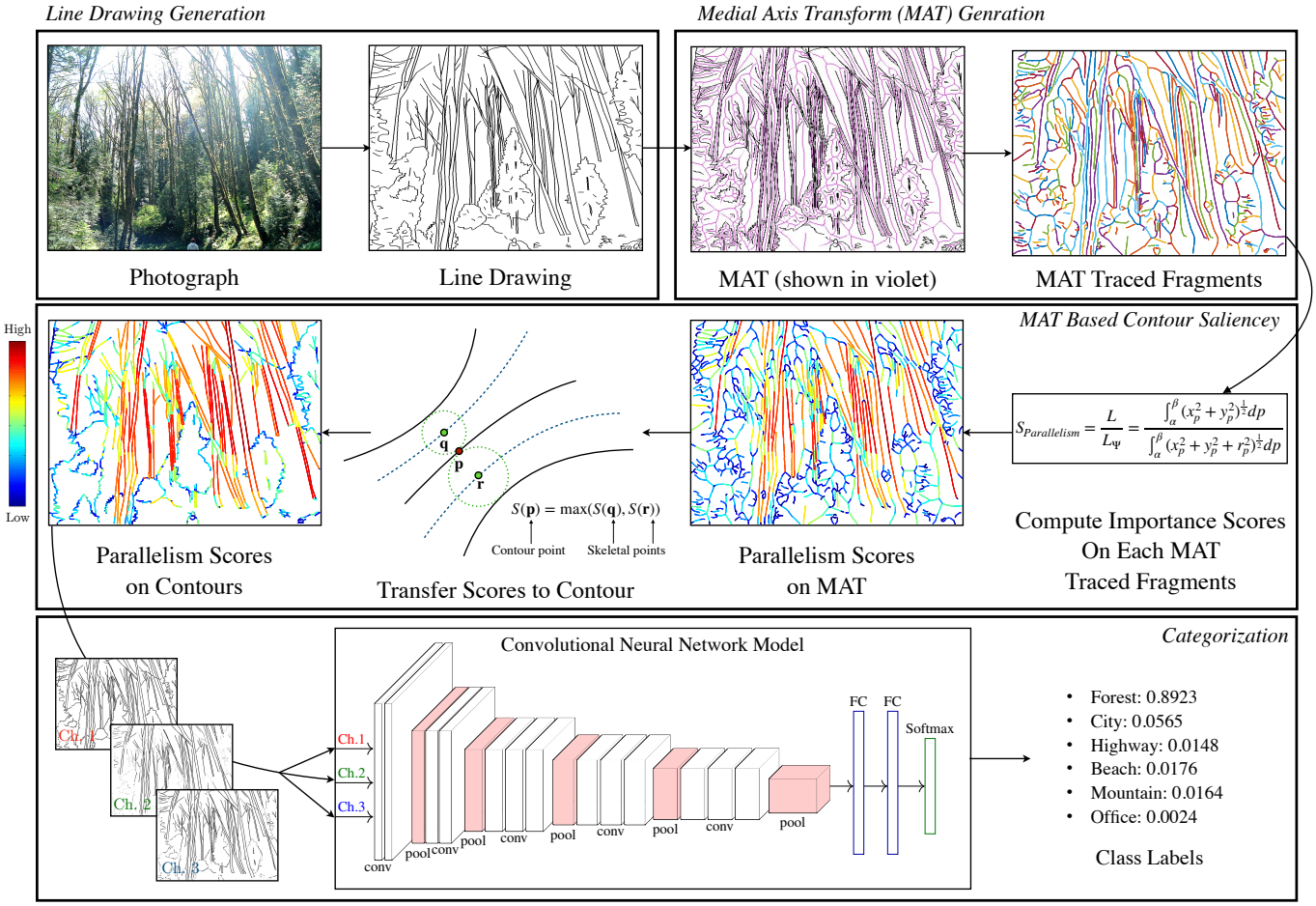
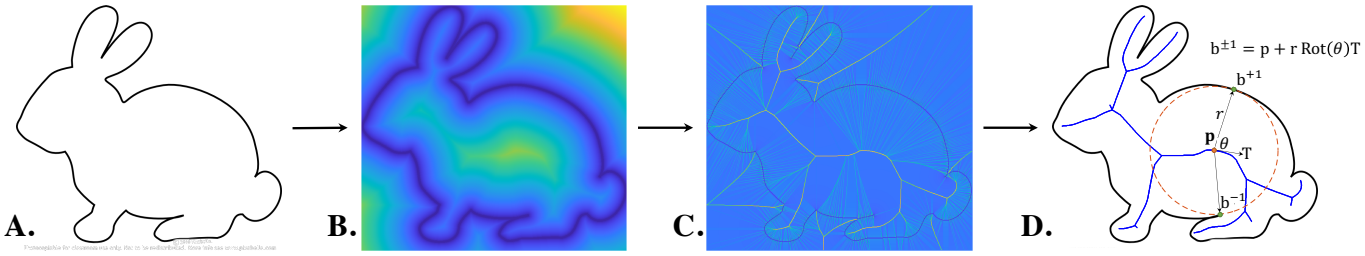


Fig. 1: (Best viewed by zooming in on the PDF.) An illustration of our approach using an example from a database of line drawings of natural scenes. In this pipeline, we first extract line drawings of the input image, then its Medial Axis Transform (MAT) is computed and then the skeletal points on the MAT are scored based upon one of our importance score measures. Finally, they are projected back onto the contours. The figure on the top left shows an example scene used in our pipeline, and the figure below presents the jet colormap visualization of its contours using our medial axis based contour importance measures, which are discussed in Section 3.

as object detection, because symmetric contours are more likely to be caused by the projection of a symmetric object than to occur accidentally. In the categorization of complex real-world scenes by human observers, local contour symmetry does indeed provide a perceptual advantage [4, 5], but the connection to the recognition of individual objects is not as straightforward as it may appear.

In vision, symmetry, proximity, good continuation, contour closure, and other cues have been used for image segmentation, curve inference, object recognition, or tasks such as object manipulation [6, 7, 8, 9]. Instantiations of such organizational principles have found their way into many computer vision algorithms and have been the subject of regular workshops on perceptual cues in artificial vision systems [10]. Inspired by Gestalt principles, [11] used measurements of edge co-occurrence statistics for the task of contour grouping in natural images. [12] found that "good continuation" can help predict the arrangement of oriented elements such as edges or line segments present in a dataset of visual scenes. [13] tried to estimate the likelihood distributions required to construct an optimal Bayesian model for contour grouping using Gestalt laws and found that each of these cues has a different power. [14] proposed a search algorithm to compute candidate closed object boundaries by combining prior probabilistic knowledge of the appearance of the object with probabilistic models using perceptual grouping.

Although scientists have studied the role of perceptual organization cues in vision for decades, these cues have not been used extensively in object recognition and scene categorization problems. This may be a result of the ability of CNN-based systems to accomplish scene categorization on challenging databases, in the presence of sufficient training data, directly from pixel intensity and colour in photographs [15, 16, 17, 18]. CNNs begin by extracting simple features, including oriented edges, which are then successively combined into more and more complex features in a succession of convolution, nonlinear activation, and pooling operations. The final levels of CNNs are typically fully connected, which enables learning of object or scene



D (Euclidean distance map to the boundary)

$\dot{\mathbf{q}} = \nabla D$ (gradient of the Euclidean distance function is computed)

$AOF = \frac{\int_{\partial R} \langle \dot{q}, N \rangle ds}{\int_{\partial R} ds}$

Final medial axis from thresholded AOF map

Fig. 2: Using the divergence theorem, [2] identifies medial axis points by considering the behaviour of the average outward flux (AOF) of the gradient of the Euclidean distance function to the boundary of a 2D object, through a shrinking disk. In particular, the limiting AOF value of all points not located on the skeleton is equal to zero. Starting with the boundary of a 2D object (**A.**), we can first compute the distance map to the boundary (**B.**), and then we can compute the AOF map (**C.**) from the distance. Finally, by keeping the non-zero AOF points (smaller than a particular threshold in the discrete space), we can get a medial axis transform (**D.**). The local geometry of a maximal inscribed disk centred at the skeletal point $\mathbf{p}$ with radius r and with object angle θ. The maximal inscribed disk touches the boundary at two points $\mathbf{b}^{\pm 1}$. As can be seen in this figure, each point on the skeleton has two or more corresponding boundary points. Therefore, given a mapping between boundary points to skeletal points, it is possible to invert that mapping to reconstruct the boundary purely from skeletal points and their properties.

categories [19, 20, 21, 22]. Unfortunately, present CNN architectures do not allow for properties of object shape to be represented explicitly. This limitation has been recognized and is the subject of some promising new work in the field [23, 24, 25, 26]. Human observers, in contrast, recognize an object's shape as an inextricable aspect of its properties, along with its category or identity [27, 28].

Comparisons between CNNs and human and monkey neurophysiology appear to indicate that CNNs replicate the entire visual hierarchy [29, 30, 31]. Does this mean that the problem of perceptual organization is now irrelevant for computer vision? In the present article, we argue that this is not the case. Rather, we show that CNN-based scene categorization systems, just like human observers, can benefit from explicitly computed contour measures derived from Gestalt-based perceptual cues. We here demonstrate the computation of these measures as well as their power to aid in the categorization of complex real-world scenes.

To effect our study, with its focus on the geometry of scene contours, we use the medial axis transform (MAT) as a representation. The MAT is defined as a set of maximal inscribed disks in the region enclosed within a boundary, along with the radii of these disks. We apply the same algorithm for computing the medial axis to analyze line drawings of scenes of increasing complexity as the one first reported in [2]. We implemented an accurate system for extraction of average outward flux skeletons, which is available here: https://github.com/mrezanejad/AOFSkeletons (see Fig. 2). This algorithm computes the average outward flux (AOF) of the gradient of the Euclidean distance function through shrinking circular disks for each point in the image plane, where the boundaries to compute the distance maps are from the line drawings of the scenes. With its explicit representation of the regions between scene contours, the medial axis allows us to directly capture importance measures related to local contour separation and local contour symmetry.

We introduce three novel measures of local symmetry using ratios of length functions derived from the medial axis radius along with skeletal segments and the curvature function derived from the medial axis tangent and the unit normal vectors. Distinct from the approach in [4], these new measures have clearer geometric interpretations and have a further advantage that they are essentially parameter-free. As ratios of commensurate quantities, these are unitless measures, which are therefore invariant to image re-sizing. We also introduce a measure of local contour separation, which is mathematically connected to the symmetry measures. We describe methods of computing our perceptually motivated importance measures from line drawings of complex real-world scenes, covering databases of increasing complexity. Fig. 1 presents an illustrative example of a photograph from an Artist Scene Database, along with all the steps applied to compute importance measures using our medial axis transform and then using it as an input channel to a CNN model. Observe how the parallelism-based measure highlights the boundaries of trees, where parallel contours in scenes are shown to facilitate categorization in scenes by humans. Our experiments will show that scene contours weighted by these measures can boost the accuracy of CNN-based categorization of scene line drawings, despite the absence of colour, texture, and shading cues. Our work shows that measures of perceptual organization cues for contours, which are simply functions of the contours themselves, are beneficial for scene categorization by computer vision methods, yet they are not automatically extracted by state-of-the-art CNN-based scene recognition systems.

In order to be able to use shape-based importance measures, we need to obtain line drawings of photographs first. We describe this step in Section 2. We then discuss how our contour importance-based measures are computed using the medial axis as a representation. We provide mathematical

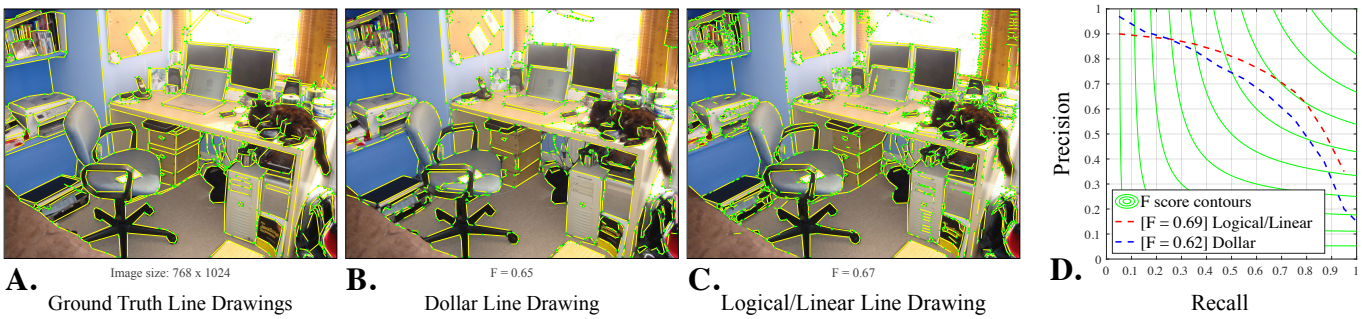


Fig. 3: A,B,C: Example scenes of different line drawing generation pipelines of this project for a qualitative comparison to the ground truth and actual photograph. D: Evaluation of line drawing generation frameworks on the Artist Scene Database. Both approaches are ranked according to their maximum F-measure $\frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$ with respect to human drawn line drawings. Iso-F curves are shown in green.

details for each of the four methods presented in this paper and show examples of how our scores vary across simple shapes in Section 3. In Section 4, we show the results from our computational and behavioural experiments. We compare human scene categorization performance with CNN performance. Moreover, we show that CNNs benefit from these additional importance measures and that these benefits are tied to these specific shape-based features. Finally we conclude with a discussion of our findings in Section 5.

2 Scene Representation by Line Drawings

Previous work has shown that, even though seemingly impoverished in their feature content, line drawings of natural scenes contain rich information about scene content [32, 33, 34, 35, 36]. This information is not necessarily distributed uniformly, rather we might find more information along with some parts of contours than others. To study this, we focus on extracting outlines from visual scenes. In this section, we show how each line drawing of a scene is generated and used in our pipeline.

2.1 Artists Generated Line Drawings

We utilize a dataset of hand-drawn tracings of scenes that helped us develop our measures and perform initial tests. Colour photographs of six categories of natural scenes (beaches, city streets, forests, highways, mountains, and offices) were downloaded from the internet, and those rated as the best exemplars of their respective categories by workers on Amazon Mechanical Turk were selected [37]. Line drawings of these photographs were generated by trained artists at the Lotus Hill Research Institute [33]. Artists traced the most important and salient lines in the photographs on a graphics tablet using a custom graphical user interface. Contours were saved as successions of anchor points. For the experiments in the present article, line drawings were rendered by connecting anchor points with straight black lines on a white background at a resolution of 1024×768 pixels. The resulting database had 475 line drawings in total with 76-80 exemplars from each of 6 categories: beaches, mountains, forests, highway scenes, city scenes, and office scenes. Images with fire or other potentially upsetting content were removed from the image set (five images total).

2.2 Machine Generated Line Drawing

Given the limited number of scene categories in the Artist Scene Database, particularly when compared to other computer vision studies, we worked to extend our results to two popular, and much larger, scene databases of photographs - MIT67 [38] (6700 images, 67 categories) and Places365 [18] (1.8 million images, 365 categories). Producing artist-generated line drawings on databases of this size was not feasible. Instead, we generated such line drawings algorithmically. We utilized two different edge detection algorithms, one from the family of learning-based edge detectors and one from the family of classical edge detectors. Each of these edge detectors produces an edge map image that represents the given image's edges, lines, or curves. Each of these edge maps was processed and traced to obtain contour fragments 1 pixel wide.

2.2.1 *Edge Detection Using Structured Forests*

Initially, in our first set of experiments, we fine-tuned the output of the Dollar edge detector [39], using the publicly available structured edge detection toolbox. From the edge map and its associated edge strength, we produced a binarized version, using per image adaptive thresholding. The binarized edge map was processed to obtain contour fragments 1 pixel wide. Each contour fragment was spatially smoothed by convolution with a Gaussian with $\sigma = 1$ pixel to mitigate discretization artifacts. The same parameters were used to produce all the MIT67 and Places365 line drawings. We confirmed that on the Artist Scene Database the machine-generated contour pixels and the artist's line drawings had 90% of the contour pixels in common. Fig. 3B shows a typical line drawing from the Artist Scene Database, produced using Dollar's framework. CNN-based scene categorization results using Dollar's edge detector have been reported in [40]. The code used to generate this type of line drawing is released here https://github.com/mrezanejad/DollarLineDrawing.

2.2.2 *Logical/Linear Operators*

Although very popular, Dollar's edge detection algorithm has some shortcomings when trying to interpret the results

as oriented edge elements, especially near contour junctions. We therefore shifted to a modified version of the Logical/Linear edge detector by [41], using their publicly available open-source implementation. This approach has the advantage of being devised to recover image curves while preserving singularities and junctions. We briefly review the edge curves as modelled in [41].

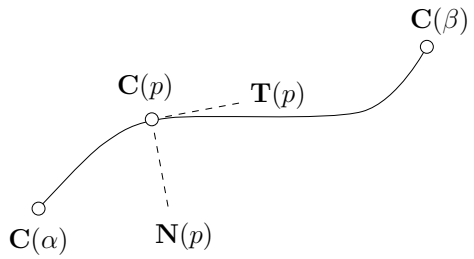


Fig. 4: An image curve $C : p \in P \rightarrow R^2$ parameterized over the interval $P = [\alpha, \beta]$ with unit tangent vector $\mathbf{T}(p)$, and unit normal vector $\mathbf{N}(p)$.

Consider an image I as an analytical intensity surface $I : R^2 \rightarrow R$, and let $\mathbf{C} : p \in P \rightarrow R^2$ represent a smooth curve parameterized by P, where P is an interval $P = [\alpha, \beta]$ in R (see Fig. 4). The normal cross section $\mathbf{N}_p(t)$ at the curve point $\mathbf{C}(p)$ is given by:

$$\mathbf{N}_p(t) = I(\mathbf{C}(p) + t\mathbf{N}(p))), \quad p \in P, \quad t \in R. \tag{1}$$

Using local structural conditions in the directions tangential and normal to the curve, the following edge curve is defined as suggested in [41]: $\mathbf{C}$ is an *Edge* iff $\mathbf{C}$ is an image curve such that the following condition holds for all $p \in P$:

$$\lim_{t \to 0^-} \mathbf{N}_p(t) > \lim_{t \to 0^+} \mathbf{N}_p(t)$$

In [41], operators are designed to respond when the edge condition is met locally in an image, and if so, an edge or a line is reported. Fig. 3C shows some typical machine-generated line drawings from the Artist Scene Database using the Logical/Linear method.

In our experiments, we produced a binarized version from the output edge map and its associated edge strength and edge directions. The implementations for this type of line drawing generation are available at https://github.com/mrezanejad/LineDrawingExtraction. We compared edges obtained using Logical/Linear operators and Dollar's edge detector with the artists' actual line drawings (see Fig. 3). In Fig 3D, we show the Precision-Recall curve for both of these methods. Logical/Linear performs slightly better in terms of F-measure performance on the Artist Scene Database.

3 Medial Axis Based Contour Importance

Owing to the continuous mapping between the medial axis and scene contours, the medial axis provides a convenient representation for designing and computing Gestalt contour importance measures based on local contour separation and local symmetry (see Fig. 2 for more on the definition of the medial axis).

We designed a measure to reflect local contour separation using the radius function along the medial axis which gives the distance to the two nearest scene contours on either side. Local parallelism between scene contours can also be directly captured by examining the degree to which the radius function along the medial axis between them remains locally constant. If contours taper off, as in the case of a set of railway tracks extending to the horizon under perspective projection, one can examine the degree to which the first derivative of the radius function is constant along a skeletal segment. Finally, if we assume that mirror symmetry is understood based on reflection on a straight symmetry axis, we can measure the curvature of the medial axis along skeletal fragments. We introduce novel measures to capture local separation, parallelism, taper, and mirror symmetry, based on these ideas. All of the introduced measures here are implemented and available for use at http://mlvtoolbox.org/ [42].

In the following we shall let p be a parameter that runs along a medial axis segment, $\mathbf{C}(p) = (x(p), y(p))$ be the coordinates of points along that segment, and $r(p)$ be the medial axis radius at each point. We shall consider the interval $p \in [\alpha, \beta]$ for a particular medial segment. The arc length of that segment is given by

$$L = \int_\alpha^\beta ||\frac{\partial \mathbf{C}}{\partial p}|| dp = \int_\alpha^\beta (x_p^2 + y_p^2)^{\frac{1}{2}} dp. \tag{2}$$

3.1 Separation

We now introduce an importance measure based on the local separation between two scene contours associated with the same medial axis segment. Consider the interval $p \in [\alpha, \beta]$. With $r(p) > 1$ in pixel units (because two scene contours cannot touch) we introduce the following contour separation-based importance measure:

$$S_{separation} = 1 - \Big(\int_\alpha^\beta \frac{1}{r(p)} dp \Big) / (\beta - \alpha). \tag{3}$$

This quantity falls in the interval $[0, 1]$. The measure increases with increasing spatial separation between the two contours. In other words, scene contours that are further apart are more salient by this measure.

3.2 Parallelism

Now consider the curve $\Psi = (x(p), y(p), r(p))$. Similar to Equation 2, the arc length of Ψ is computed as:

$$L_\Psi = \int_\alpha^\beta ||\frac{\partial \Psi}{\partial p}|| dp = \int_\alpha^\beta (x_p^2 + y_p^2 + r_p^2)^{\frac{1}{2}} dp. \tag{4}$$

When two scene contours are close to being parallel locally, $r(p)$ will vary slowly along the medial segment. This motivates the following parallelism importance measure:

$$S_{parallelism} = \frac{L}{L_\Psi} = \frac{\int_\alpha^\beta (x_p^2 + y_p^2)^{\frac{1}{2}} dp}{\int_\alpha^\beta (x_p^2 + y_p^2 + r_p^2)^{\frac{1}{2}} dp}. \tag{5}$$

This quantity also falls in the interval $[0, 1]$ and is invariant to image scaling since the integral involves a ratio of unitless quantities

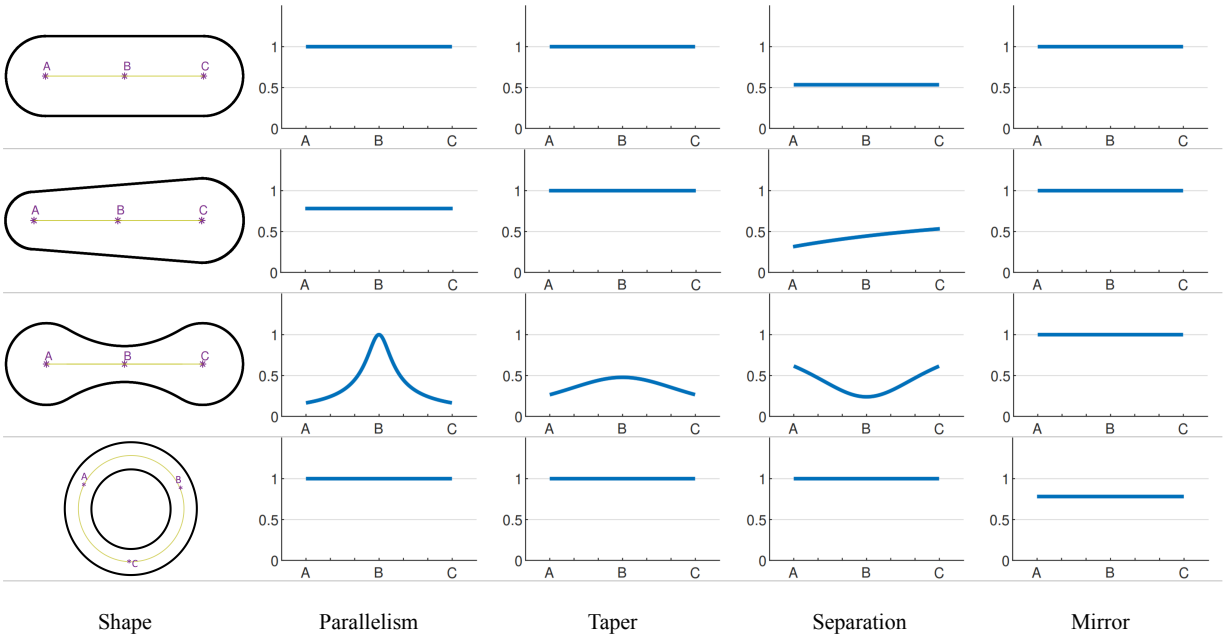


Fig. 5: An illustration of parallelism score, taper score, contour separation score, and mirror symmetry score for four different contour configurations. See text for a discussion.

3.3 Taper

A notion that is closely related to that of parallelism is taper; two scene contours are taper symmetric when the medial axis between them has a radius function that is changing at a constant rate, such as the edges of two parallel contours receding into depth when viewed in perspective. To capture this notion of symmetry, we introduce a slight variation where we consider a type of arc-length of a curve $\Psi' = (x(p), y(p), \frac{dr(p)}{dp})$. Specifically, we introduce the following taper importance measure:

$$S_{Taper} = \frac{L}{L_{\Psi'}} = \frac{\int_\alpha^\beta (x_p^2 + y_p^2)^{\frac{1}{2}} dp}{\int_\alpha^\beta (x_p^2 + y_p^2 + (r r_{pp})^2)^{\frac{1}{2}} dp}. \quad (6)$$

The bottom integral is not exactly an arc-length, due to the multiplication of r_{pp} by the factor r. This modification is necessary to make the overall ratio unitless. This quantity also falls in the interval $[0, 1]$ and is invariant to image scaling. The measure is designed to increase as the scene contours on either side become more taper symmetric, as in the shape of a funnel, or the sides of a railway track receding in the distance.

3.4 Mirror Symmetry

Mirror Symmetry is commonly thought to be the symmetry caused by reflection across a straight axis. Many objects and shapes we see in nature are mirror-symmetric, making mirror symmetry a useful heuristic for detecting likely object locations. In this section, we propose a notion of local mirror symmetry based on the curvature of the medial axis. A straight medial axis allows for the contours on either side to be related to each other by reflection over that straight axis. A strongly curved medial axis does not allow for the same mirror symmetry. To measure curvature of the medial axis, we consider Euclidean curvature which is defined in terms of arc-length parameterization. For a smooth medial axis segment $\mathbf{C}(p) = (x(p), y(p))$, the unit tangent is defined as $\mathbf{T}(p) = \frac{(x_p, y_p)}{\sqrt{x_p^2 + y_p^2}}$ and the unit normal is defined as $\mathbf{N}(p) = \frac{(-y_p, x_p)}{\sqrt{x_p^2 + y_p^2}}$. The Euclidean curvature $\kappa(p)$ can be represented as:

$$\kappa(p)\mathbf{N}(p) = \frac{1}{\sqrt{x_p^2 + y_p^2}} \frac{\partial}{\partial p} \left(\frac{(x_p, y_p)}{\sqrt{x_p^2 + y_p^2}} \right). \quad (7)$$

As the radius of curvature is inversely proportional to "straightness", the mirror symmetry peaks when the radius of curvature goes to infinity. Thus, we consider the inverse of the radius of curvature as the local mirror symmetry score:

$$S_{Mirror} = \frac{\int_\alpha^\beta R_{curv}(p)}{\beta - \alpha} = \Big(\int_\alpha^\beta \frac{1}{\kappa(p)(\beta - \alpha)} dp \Big). \quad (8)$$

Unlike parallelism and taper, this measure is not scale invariant. To gain an intuition behind these perceptually driven contour importance measures, we provide four illustrative examples in Fig. 5. The measures are not computed point-wise, but rather for a small interval $[\alpha, \beta]$ centered at each medial axis point (see Section 4.1 for details). When the contours are parallel, all four measures are constant along the medial axis (first row). The second row has high taper and mirror symmetry but comparably lower parallelism, with contour separation score increasing from left to right. For the dumbbell shape, the first three measures vary (third row), while mirror symmetry gives a constant score of 1 from point A to C. Finally, for the ring shape, the first three measures are equal to one, while mirror symmetry receives a score less than one as the medial axis bends at a fixed

rate. Fig. 6 shows different examples of scene line drawings from the Artist Scene Database weighted by our perceptual importance measures.

4 Scene Categorization Experiments and Results

4.1 Computing Contour Score

Computing contour scores for each line drawing required a number of steps. First, each connected region between scene contours was extracted. Second, we computed an AOF map for each of these connected components using a disk of radius 1 pixel, with 60 discrete sample points on it. We used a threshold of $\tau = 0.25$ on the AOF map, which corresponds to an object angle $\theta \approx 23$ degrees [2] (please see Appendix I), to extract skeletal points. A typical example appears in Fig. 1 (middle left). The resulting AOF skeleton was partitioned into medial curves between branch points or between a branch point and an endpoint. We computed a discrete version of each of the three importance measures in Section 3, within an interval $[\alpha, \beta]$ of length $2K + 1$, centred at each medial axis point, with $K = 5$ pixels.

Each scene contour is associated with two medial curves on either side. Therefore, each scene contour point receives one score value from each side. In the present framework, we pick the maximum of the two values for each point as illustrated in Fig. 1 (middle row).

4.2 50-50 Splits of Contour Scenes

In the following of this section, we created stimuli that are either the same as the intact line drawing or contain only 50% pixels of the pixels of the intact line drawing. Accordingly, we created splits of the higher 50% and the lower 50% of the contour pixels in each image of the Artist Scene Database and MIT67 data sets, using the four importance measures, parallelism, taper, mirror symmetry and local contour separation. An example of the original intact line drawing and each of the four sets of splits is shown in Fig. 7, for an office scene from the Artist Scene Database.

4.3 Human Experiments on 50-50 Splits of Contour Scenes

Our first set of scene categorization experiments is motivated by our earlier work that shows that human observers benefit from contour symmetry in scene recognition from contours [4]. Our goal is to examine whether a CNN-based system also benefits from such perceptually motivated cues.

On the Artist Scene Database, human observers were tasked with determining to which of six scene categories an exemplar belonged. The observers were psychology undergraduates in an introductory psychology course at the University of Toronto. The study was approved by the Research Ethics Board of the University of Toronto, and observers gave informed consent prior to participation. In order to keep the number of conditions in the experiments manageable, the four different types of importance measures were tested in separate experiments. For the parallelism and taper importance measure images, there were 29 observers (21 Female, 8 Male, mean age $= 23.0$, range 20 to 38). For the separation importance measures, a new set of 23 observers participated (14 Female, 9 Male, mean age $= 19.1$, range 18 to 23). For the mirror symmetry importance measure, a new set of 28 observers participated (18 Female, 9 Male, 1 Other, mean age $= 19.1$, range 18 to 23).

Due to the COVID-19 pandemic, there were changes to the data collection procedure. Different sets of participants needed to be used, and the mirror symmetry portion of the study was changed from an in-lab to an online format. Prior to collecting the mirror symmetry data, a pilot study verified that the in-lab parallelism results would replicate in the online experiment. All sets of observers were shown either the artist-generated line drawing or the high 50% or the low 50% splits by one of the importance measures. Images were presented for only 58 ms and were followed by a perceptual mask, making the task difficult for observers, who would otherwise perform near 100% correct. For the online version of the study, the stimulus duration needed to be increased to 159 ms in order for performance on the intact images to be roughly equivalent to the intact performance for the other datasets. The results with this short image presentation duration, shown in Fig. 8 (left), demonstrate that human performance is consistently better with the high 50% (more salient) than the low 50% condition, for each importance measure. The performance was slightly higher for all conditions when testing the separation splits.

4.4 CNN-based Experiments on 50-50 Splits of Contour Scenes

Carrying out CNN-based recognition on the Artist Scene Database and MIT67 line drawing datasets presents the challenge that they are too small to train a large model, such as VGG-16, from scratch. To the best of our knowledge, no CNN-based scene categorization work has so far focused on line drawings of natural images. We, therefore, use CNNs that are pre-trained on RGB photographs for our experiments.

For our experiments on the Artist Scene Database and MIT67 datasets (using Dollar's edge detector [39] for machine-generated line drawings), we use the VGG16 convolutional layer network architecture [43] with weights pre-trained on ImageNet. The last three layers of the VGG16 network used for fine-tuning are replaced with two fully connected layers and a softmax layer, where the output label is one of the categories in each of our datasets. The images are processed by this network, and the final classification layer produces an output vector in which the top-scoring index is selected as the prediction output.

For all experiments on the Artist Scene Database, we use 5-fold cross-validation. Top-1 classification accuracy is given, as a mean over the 5 folds, in Fig. 8 (left - middle). The CNN-based system mimics the trend we saw in human observers, namely that performance is consistently better for images that contained the highest 50% contour pixels according to each of the Gestalt-based measures we propose in this paper. We interpret this as evidence that all of those Gestalt-motivated importance measures are beneficial for scene categorization in both computer and human vision.

For MIT67 we use the provided training/test splits and present the average results over 5 trials. The CNN-based

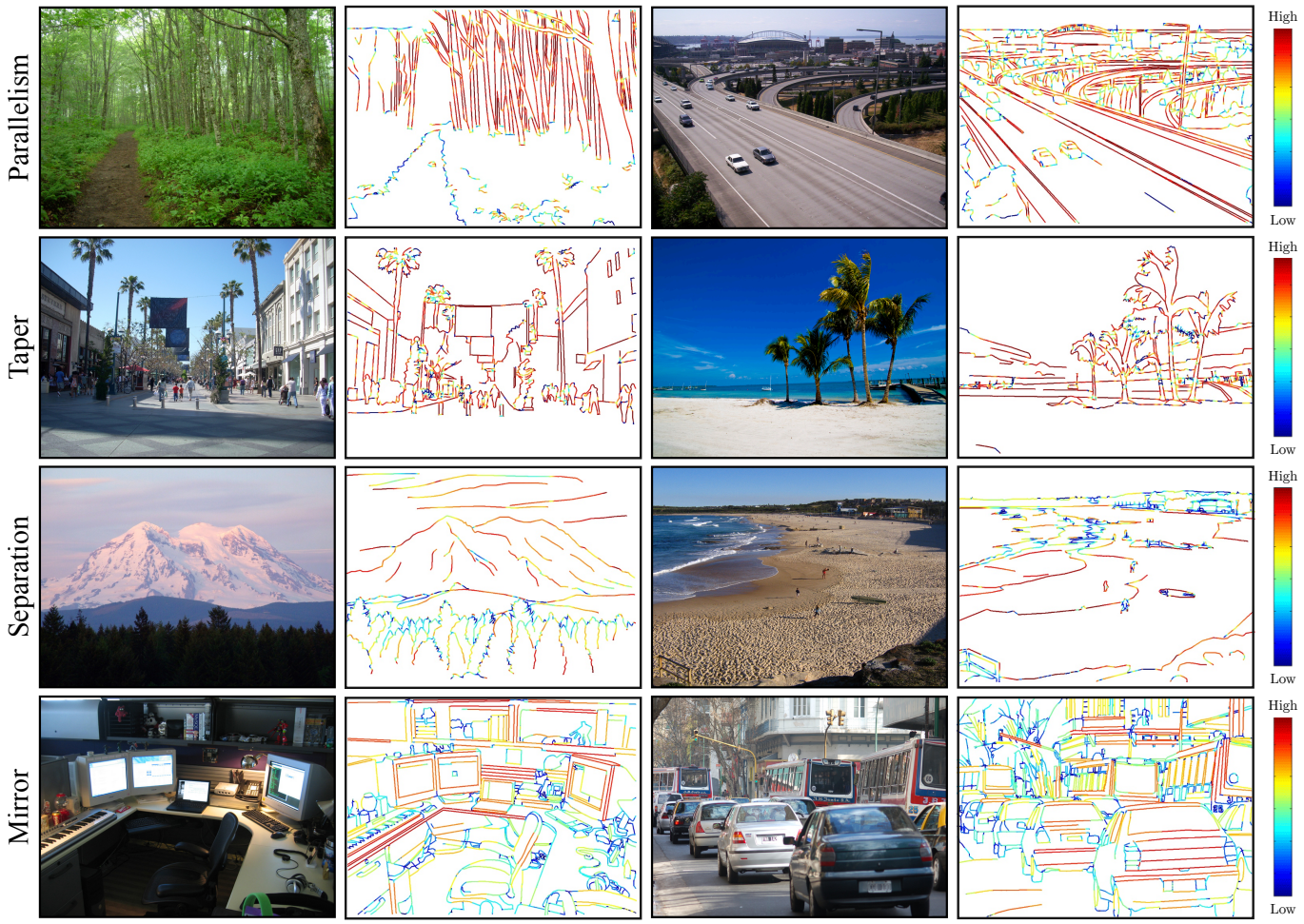


Fig. 6: (Best viewed by zooming in on the PDF.) Examples of original photographs and the corresponding mirror symmetry, separation, parallelism, and taper scores on scene contours, using a jet colormap to show increasing values.

categorization results are shown in Fig. 8 (right). It is striking that even for this more challenging database, the CNN-based system still follows the trend we saw in human observers, i.e., that performance is better for the high 50% split than for the low 50% split of each of the four measures and is well above chance. It is worth noting that neither humans nor CNNs were trained on splits, but only shown splits at test time.

4.5 Experiments with Score Weighted Contours

While we would expect that network performance would degrade when losing half the input pixels, the splits also reveal a significant bias in favour of our importance measures to support scene categorization. Can we exploit this bias to improve network performance when given intact contours? To address this question, we carried out a second experiment where we explicitly encoded importance measures for the CNN by feeding different features into the input channels of the pre-trained network. To this end, we added channels with contours weighted by the importance measures to the input to the CNNs, as illustrated in Fig. 1 (bottom subfigure). These contour importance images replace the standard three-channel (R, G, B) inputs to the network. We chose to include only the three input channels in the model design to avoid making significant modifications to the architecture, such as introducing new convolution layers with additional weights. Our goal was to determine whether the data provided to the network could influence the results. Essentially, we aimed to examine whether the network could leverage shape-based cues in an apples-to-apples comparison. For all experiments, training was done on the feature maps generated by the new feature-coded images.

We conducted various experiments to determine whether CNN models, specifically VGG16, benefit from the additional information provided by computing shape-based measures. We employed two training schemes, depending on the dataset size of the experiments conducted: (a) fine-tuning a small set of layers for the Artist Scene and MIT67 datasets, and (b) fine-tuning weights across all layers for the much larger Places365 dataset.

4.5.1 Experiments on Artist Scene and MIT67

In the experiments conducted on Artist Scene and MIT67 databases, we froze all layers of VGG16, except for the last two fully-connected layers of the pre-trained networks. We fine-tuned them using our feature-coded inputs, training on the feature maps they provided. We used Logical/Linear

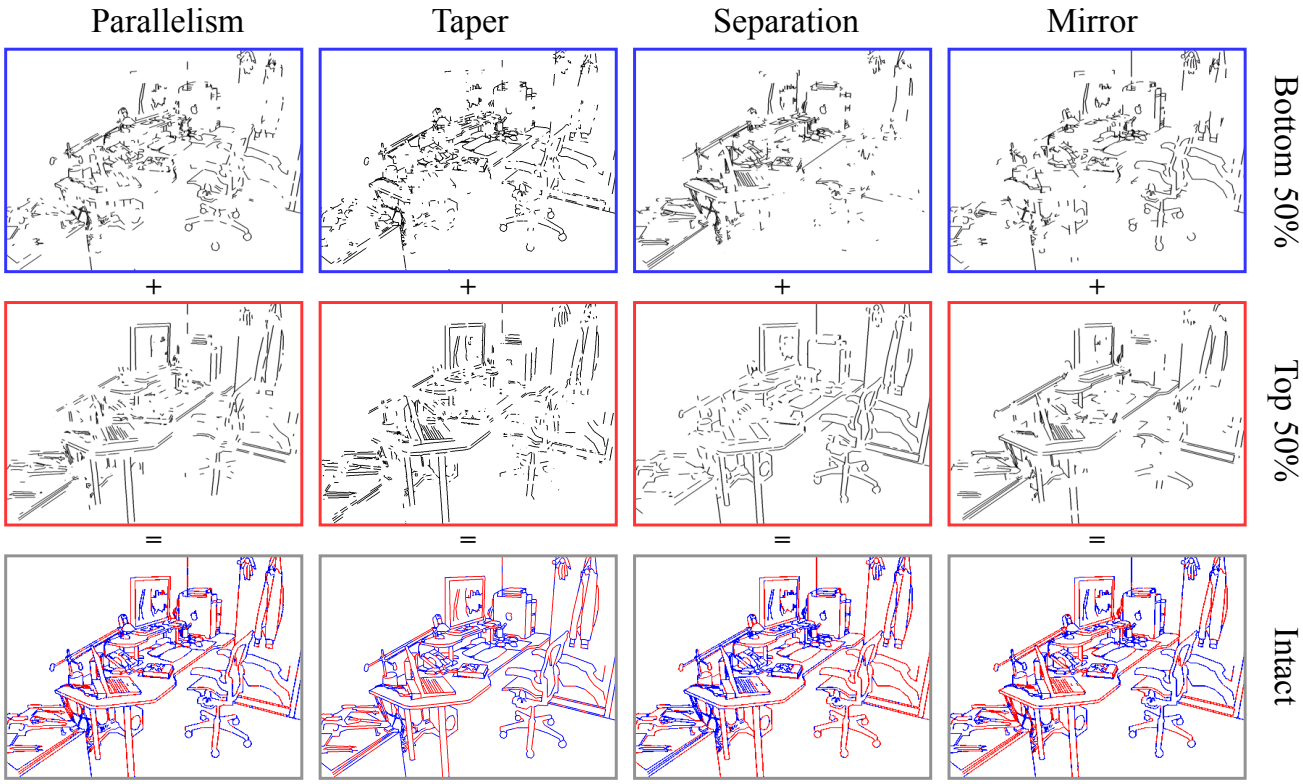


Fig. 7: We consider an office scene and create splits of the artist generated line drawings, each of which contains 50% of the original pixels, based on local contour separation (first column), mirror symmetry (second column), parallelism (third column), and taper (fourth column) based importance measures.

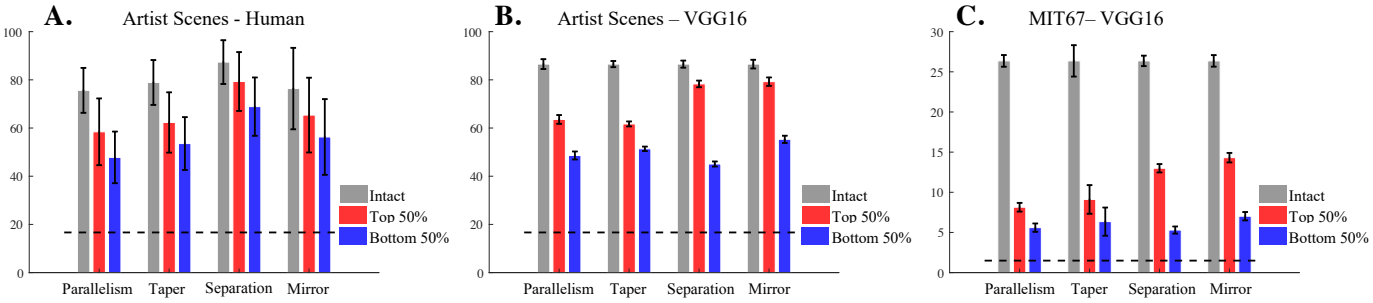


Fig. 8: A comparison of human scene categorization performance (**A**) with CNN performance (**B,C**). As with the human observer data, CNNs perform better on the high 50% split according to each importance measure, than the low 50% split. In each plot, chance level performance (1/6 for Artist Scene Database and 1/67 for MIT67) is shown with a dashed line.

operators from [41] to generate line drawings for the Artist Scene Database and MIT67. The results are presented in Table 1. Due to the limited size of these databases, this approach was employed to obtain meaningful results. As mentioned earlier, we initially used Dollar's edge detector output [39] and those results are reported for the Artist Scene Database and for MIT67 in Table 1. We then repeated the same set of experiments with our updated line drawings using Logical/Linear operators [41]. This time, we used VGG16-H (pre-trained on both Imagenet and Places365 [18]) in addition to VGG16 (pre-trained on Imagenet).

First, it is noticeable that the Logical/Linear edge detection framework gives better results than Dollar's edge detection algorithm, presumably because of the importance of singularities and junctions for scene categorization. Second, it is apparent that performance is consistently boosted by adding the importance-weighted contour channels, independent of which machine-generated line drawing algorithm is used. In all cases, the largest performance boost comes from a combination of contours, mirror symmetry and separation scores. We believe this is because mirror symmetry is conceptually the most similar cue to what people generally understand as symmetry. The local separation score, on the other hand, provides a more distinct and complementary perceptual cue.

Channel Setp			Artist	MIT67 (Logical/Linear)	MIT67 (Dollar)
Ch.1	Ch.2	Ch.3	VGG16	VGG16	VGG16
Photos (RGB)			99.62	79.49	79.49
Contour	Contour	Contour	92.50	60.29	55.34
Contour	Contour	Parallelism	94.16	61.05	54.98
Contour	Contour	Taper	95.06	63.31	56.52
Contour	Contour	Separation	96.56	62.92	55.09
Contour	Contour	Mirror	97.02	63.72	58.06
Contour	Parallelism	Taper	96.61	62.88	57.01
Contour	Parallelism	Mirror	97.26	63.70	57.50
Contour	Parallelism	Separation	98.40	64.25	57.83
Contour	Taper	Mirror	97.77	63.51	59.86
Contour	Taper	Separation	97.93	65.79	58.32
Contour	Separation	Mirror	**98.99**	**67.28**	**60.65**
Parallelism	Taper	Separation	95.96	63.48	56.72
Parallelism	Taper	Mirror	96.42	63.87	57.13
Parallelism	Separation	Mirror	97.34	64.96	58.80
Taper	Separation	Mirror	97.20	64.09	59.93
Randomized Condition			92.37	60.39	56.81
Local Intensity			92.05	60.19	55.31
Magnitude of Gradient			91.10	59.35	56.94
Magnitude of Curvature			92.22	60.69	56.32
*Maximum gain over contours			**6.49**	**6.99**	**5.31**

TABLE 1: Top 1 level performance in a 3-channel configuration, on the Artist Scene and MIT67 databases, with fine-tuning. TOP ROW: Results of the traditional R,G,B input configuration where the original photos are used. OTHER ROWS: Combinations of intact scene contours, and scene contours weighted by our importance measures, where each letter stands for a specific input channel. *Denotes the maximum gains observed by adding additional score channels to the contours. Contour, Parallelism, and Separ. are short abbreviations for Contours, Parallelism, and Separation channels.

Channel Setp			**Top 1**	**Top 5**
Ch.1	Ch.2	Ch.3	VGG-16	
Photos (RGB)			53.76	83.62
Contour	Contour	Contour	40.32	69.03
Contour	Contour	Parallelism	41.93	70.05
Contour	Contour	Taper	42.06	71.15
Contour	Contour	Separation	42.54	71.19
Contour	Contour	Mirror	43.07	71.88
Contour	Parallelism	Taper	42.89	72.33
Contour	Parallelism	Mirror	44.02	76.95
Contour	Parallelism	Separation	44.67	76.73
Contour	Taper	Mirror	44.77	76.81
Contour	Taper	Separation	43.86	77.40
Contour	Separation	Mirror	**45.06**	**78.14**
Parallelism	Taper	Separation	43.98	76.27
Parallelism	Taper	Mirror	43.72	75.47
Parallelism	Separation	Mirror	43.44	76.39
Taper	Separation	Mirror	44.03	75.20
Randomized Condition			39.08	69.67
Local Intensity			39.64	67.35
Mag. of Gradient			40.19	68.54
Mag. of Curvature			38.73	68.56
*Gain over contours			**+4.74**	**+9.11**

TABLE 2: Places365 (VGG16) - Top 1 and Top 5 performances in a 3-channel configuration on Places365 (see text). In each case, all layers of the network were trained from scratch. The top row shows the results of the traditional R,G,B input configuration, while the others show combinations of scene contours and scene contours weighted by our importance measures. Here, the machine generated MIT67 line drawings are based on the Logical/Linear edge detection framework [41]. *Denotes the maximum gains observed by adding additional score channels to the contours. Please note that the channels here are set up similar to the Table 1.

4.5.2 Control Conditions

For MIT67 the performance of 79.49% on photographs is consistent with that reported in [18]. Remarkably, 75% of this level of performance (a level of 60.73%) can be obtained by using *only* Logical/Linear line drawings. To analyze the effect of adding additional features (i. e. importance-based measures), we implemented a new set of control experiments, where we looked at the effect of feeding the neural network with these extra channels. The first test included randomizing the importance measure across the contours. For each line drawing of a scene, we considered the set of contour fragments:

$$LineDrawing = \{C_1, C_2, ..., C_n\}$$

where $C_i = \{(x_{i1}, y_{i1}), ..., (x_{im}, y_{im})\}$ represents a contour fragment. We also consider a specific channel of importance measure

$$SalienceChannel = \{S_1, ..., S_n\}$$

for that line drawing where $S_i = \{s_{i1}, ..., s_{im}\}$ is the importance scores corresponding to C_i. To create a randomized condition, we take the set of $SalienceChannel$ and randomly permute S_is. The random permutation results in an unequal length of S_is for C_is. We then take the S_is and linearly downsample or upsample for its corresponding C_i so all the contour fragments on each C_i get an equivalent random score. Results of the "Randomized Condition" are reported in Table 1.

In addition, we provide further control conditions using "local intensity", "the magnitude of the gradient", and "the magnitude of curvature" at each pixel of a contour fragment as additional input to the neural network. None of the four control conditions improved categorization accuracy above the accuracy for contours alone (Table 1). These results

show that the added performance obtained by adding score channels is **not accidental** and the added channels are providing additional information that the network is not already exploring.

4.5.3 Experiments on Places365

For the smaller Artist Scene and MIT67 datasets, we used a pretrained model with frozen weights, except for the last three layers, whose weights were fine-tuned. In order to be able to draw definitive conclusions on the benefits of MAT-based importance weighted contour scores for scene categorization, one would wish to train all layers of the network using such inputs. This requires a much bigger dataset than either of the two previously used sets. We therefore chose the much larger Places365 dataset, which contains 1.8 million images (referenced by http://places2.csail.mit.edu/download.html).

We experimented with training the entire VGG16 model, layer by layer, in several ways, including training with all weights initialized randomly. Regardless of our choice of starting weights, all training procedures led to very similar results. In the present article, we report the scores that we obtained by training the network fully with initial weights obtained from pretraining on ImageNet, which is standard practice in the field. For each experiment, we either retrained the network using contours or one of the combinations of importance-weighted contours, as shown in Tables 1 and 2.

For the Places365 dataset, chance recognition performance would be at 1/365 or 0.27%. Our results using the Logical/Linear method for machine-generated line drawings are shown in Table 2. For the training tasks in Tables 1 and 2, we used the PyTorch script provided by https://github.com/pytorch/examples/tree/master/imagenet. We ran each experiment for 100 epochs with a learning rate of 0.001 and default patience of 10 epochs (number of epochs with no improvement after which learning rate will be reduced).

Once again we see a clear and consistent trend of a benefit using importance-weighted contours as additional feature channels to the contours themselves, with the best performance gain coming from the addition of parallelism and separation scores. Finally, note that in the Artist Scene, MIT67 and Places365 databases, the percentage of contour ink pixels over all the RGB pixels in the photographs, is only 7.44%, 8.75% and 8.32%, on average. Therefore, recognition based on contours alone is more efficient than that based on photographs, in the sense of sparsity in the input data.

5 Discussion

In this article, we have demonstrated the importance of shape-based visual cues in the context of perception, specifically the task of scene categorization. We have shown that scene contours, weighted by perceptually motivated contour importance measures, can boost CNN-based scene categorization accuracy, despite the absence of colour, texture, and shading cues. Our experiments reveal that measures of perceptual cues for contours, which are simply functions of the contours themselves, are beneficial for scene categorization by computers. The critical remaining question is whether this omission is due to the CNN architecture being unable to model these weights or whether this has to do with the (relatively standard) training regime. We leave this question for further study. Scene categorization performance based on contours can surpass 80% of the best reported results on the underlying photographs. Whereas this shape information is reflected in the images themselves, it does not appear to be directly learned by present state-of-the-art CNN-based scene recognition systems. The results obtained by our approach are in line with recent work by Geirhos et al. [23, 24], suggesting that using shape by CNNs for recognition/categorization is a promising direction for future work.

The measures we have introduced have also been used to separate the contour pixels in a given scene into the more salient and the less salient halves. We have demonstrated that human observers are better at categorizing scenes containing only the more salient halves in a rapid-categorization psychophysical experiment in Section 4.3. More interestingly, we show that CNNs exhibit the same trend which inspired us to ask this question, do perceptually motivated line drawing-based importance measures also aid scene categorization in machine vision? We investigated the answer to this question by using the weighted contours in a series of exhaustive training tasks where CNNs were fed contours with these perceptually weighted scores. This way, we evaluated the methods we developed for contour-based scene abstraction and categorization, we also significantly extended the contour databases previously used for benchmarking computer vision systems used for such tasks. Notably, existing databases of line drawings contain hundreds or thousands of images. We generated a database containing millions of line drawings, created from photographs of complex scenes in the Places365 dataset [18, 38] by using the Logical/Linear edge detection framework of Iverson and Zucker [41]. Our work demonstrates the promise of perceptual organization principles from human vision in improving the capabilities of computer vision-based scene categorization and recognition systems.

In recent work, Geirhos et al. [23] showed that CNNs that are used to recognize object classes are biased towards learning texture in visual inputs, rather than complex representations of object shapes (also see [26]). Our present effort in scene categorization from line drawings demonstrates that shape and contour geometry are not exploited by patch-based CNN systems, complementing and resonating with the ideas of [23]. Here, we have presented results to showcase the importance of contours and perceptually weighted contours. Looking forward, it might be useful to identify and formulate more general principles to guide the perceptual organization of local contour elements, providing support to the human visual system for the understanding of cluttered, real-world scenes in a more comprehensive way.

Future directions for this work include identifying and formulating a complete list of principles that guide the perceptual organization of local contour elements, providing support to the human visual system in the understanding of cluttered, real-world scenes in a more comprehensive way. This paper provides a path to such a goal which includes describing shapes using medial axis representation.

Previous work has showcased the strength of extracting a medial representation from colour photographs of objects [44, 45, 46, 47, 48, 49]. These ideas can be extended to work well with scenes. Connecting the representation to the cues computed here can give us a trained end-to-end system that is likely capable of performing various visual tasks better with much less training/tuning. Extending these ideas and assessing the viability of these principles for facilitating the categorization of complex real-world environments in computer vision systems is a key next step to consider. Interpreting the role of perceptual organization cues in tasks such as recognition or categorization of objects or entire scenes in biological and artificial vision systems are also promising directions for future work.

ACKNOWLEDGMENTS

We are grateful to the University of Toronto, NSERC, Samsung, and Sony for research support.

REFERENCES

[1] J. Goodnow, *Children's Drawing (The Developing Child)*. HarperCollins Publishers, 1980.

[2] P. Dimitrov, J. N. Damon, and K. Siddiqi, "Flux invariants for shape," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 1. IEEE, 2003, pp. I–835.

[3] K. Koffka, "Perception: An introduction to the Gestalt-theorie," *Psychological Bulletin*, vol. 19, no. 10, pp. 531–585, 1922.

[4] J. Wilder, M. Rezanejad, S. Dickinson, K. Siddiqi, A. Jepson, and D. B. Walther, "Local contour symmetry facilitates scene categorization," *Cognition*, vol. 182, pp. 307 – 317, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0010027718302506

[5] ——, "Neural correlates of local parallelism during naturalistic vision," *Plos one*, vol. 17, no. 1, p. e0260266, 2022.

[6] D. Marr and H. K. Nishihara, "Representation and recognition of the spatial organization of three-dimensional shapes," *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 200, no. 1140, pp. 269–294, 1978.

[7] I. Biederman, "Recognition-by-components: a theory of human image understanding." *Psychological Review*, vol. 94, no. 2, p. 115, 1987.

[8] J. H. Elder and S. W. Zucker, "Computing contour closure," in *European Conference on Computer Vision*. Springer, 1996, pp. 399–412.

[9] S. Sarkar and K. L. Boyer, "Perceptual organization in computer vision: status, challenges, and potential," *Computer Vision and Image Understanding*, vol. 76, no. 1, pp. 1–5, 1999.

[10] I. Sofer, S. M. Crouzet, and T. Serre, "Explaining the timing of natural scene understanding with a computational model of perceptual categorization," *PLOS Computational Biology*, vol. 11, no. 9, pp. 1–20, 09 2015. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1004456

[11] W. S. Geisler, J. S. Perry, B. Super, and D. Gallogly, "Edge co-occurrence in natural images predicts contour grouping performance," *Vision Research*, vol. 41, no. 6, pp. 711–724, 2001.

[12] M. Sigman, G. A. Cecchi, C. D. Gilbert, and M. O. Magnasco, "On a common circle: Natural scenes and gestalt rules," *Proceedings of the National Academy of Sciences*, vol. 98, no. 4, pp. 1935–1940, 2001. [Online]. Available: https://www.pnas.org/content/98/4/1935

[13] J. H. Elder and R. M. Goldberg, "Ecological statistics of Gestalt laws for the perceptual organization of contours," *Journal of Vision*, vol. 2, no. 4, pp. 5–5, 08 2002. [Online]. Available: https://doi.org/10.1167/2.4.5

[14] J. Elder, A. Krupnik, and L. Johnston, "Contour grouping with prior models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 6, pp. 661–674, 2003.

[15] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 806–813.

[16] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[17] S. Hoo-Chang, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, p. 1285, 2016.

[18] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2018.

[19] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun RGB-D: A RGB-D scene understanding benchmark suite," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 567–576.

[20] S. Bai, "Growing random forest on deep convolutional neural networks for scene categorization," *Expert Systems with Applications*, vol. 71, pp. 279–287, 2017.

[21] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[23] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bygh9j09KX

[24] R. Geirhos, C. R. M. Temme, J. Rauber, H. H. Schütt,

M. Bethge, and F. A. Wichmann, "Generalisation in humans and deep neural networks," in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 7538–7550.

[25] N. Baker, H. Lu, G. Erlikhman, and P. J. Kellman, "Deep convolutional networks do not classify based on global object shape," *PLOS Computational Biology*, vol. 14, no. 12, pp. 1–43, 12 2018. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1006613

[26] B. Carter, S. Jain, J. W. Mueller, and D. Gifford, "Overinterpretation reveals image classification model pathologies," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[27] P. J. Kellman and T. F. Shipley, "A theory of visual interpolation in object perception," *Cognitive Psychology*, vol. 23, no. 2, pp. 141–221, 1991.

[28] E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem, "Basic objects in natural categories," *Cognitive Psychology*, vol. 8, no. 3, pp. 382 – 439, 1976. [Online]. Available: http://www.sciencedirect.com/science/article/pii/001002857690013X

[29] U. Güçlü and M. A. J. van Gerven, "Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream," *Journal of Neuroscience*, vol. 35, no. 27, pp. 10 005–10 014, 2015. [Online]. Available: http://www.jneurosci.org/content/35/27/10005

[30] C. F. Cadieu, H. Hong, D. L. Yamins, N. Pinto, D. Ardila, E. A. Solomon, N. J. Majaj, and J. J. DiCarlo, "Deep neural networks rival the representation of primate IT cortex for core visual object recognition," *PLoS Computational Biology*, vol. 10, no. 12, p. e1003963, 2014. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1003963

[31] M. Schrimpf, J. Kubilius, M. J. Lee, N. A. R. Murty, R. Ajemian, and J. J. DiCarlo, "Integrative benchmarking to advance neurally mechanistic models of human intelligence," *Neuron*, 2020. [Online]. Available: https://doi.org/10.1016/j.neuron.2020.07.040

[32] I. Biederman, *On the semantics of a glance at a scene*, 1981.

[33] D. B. Walther, B. Chai, E. Caddigan, D. M. Beck, and L. Fei-Fei, "Simple line drawings suffice for functional mri decoding of natural scene categories," *Proceedings of the National Academy of Sciences*, vol. 108, no. 23, pp. 9661–9666, 2011.

[34] D. B. Walther and D. Shen, "Nonaccidental properties underlie human categorization of complex natural scenes," *Psychological science*, p. 0956797613512662, 2014.

[35] J. E. Fan, R. D. Hawkins, M. Wu, and N. D. Goodman, "Pragmatic inference and visual abstraction enable contextual flexibility during visual communication," *Computational Brain & Behavior*, vol. 3, no. 1, pp. 86–101, 2020.

[36] C. Zou, Q. Yu, R. Du, H. Mo, Y. SONG, T. Xiang, C. Gao, B. Chen, H. Zhang *et al.*, "Sketchyscene: Richly-annotated scene sketches." European Conference on Computer Vision, 2018.

[37] A. Torralbo, D. B. Walther, B. Chai, E. Caddigan, L. Fei-Fei, and D. M. Beck, "Good exemplars of natural scene categories elicit clearer patterns than bad exemplars but not greater bold activity," *PloS one*, vol. 8, no. 3, p. e58594, 2013.

[38] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 413–420.

[39] P. Dollár and C. L. Zitnick, "Structured forests for fast edge detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1841–1848.

[40] M. Rezanejad, G. Downs, J. Wilder, D. B. Walther, A. Jepson, S. Dickinson, and K. Siddiqi, "Scene categorization from contours: Medial axis based salience measures," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[41] L. A. Iverson and S. W. Zucker, "Logical/linear operators for image curves," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 10, pp. 982–996, 1995.

[42] D. B. Walther, D. Farzanfar, S. Han, and M. Rezanejad, "The mid-level vision toolbox for computing structural properties of real-world images," *Frontiers in Psychology*, vol. 14, p. 1322.

[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015. [Online]. Available: http://arxiv.org/abs/1409.1556

[44] T. Sie Ho Lee, S. Fidler, and S. Dickinson, "Detecting curved symmetric parts using a deformable disc model," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 1753–1760.

[45] W. Shen, K. Zhao, Y. Jiang, Y. Wang, X. Bai, and A. Yuille, "Deepskeleton: Learning multi-task scale-associated deep side outputs for object skeleton extraction in natural images," *IEEE Transactions on Image Processing*, vol. 26, no. 11, pp. 5298–5311, 2017.

[46] Y. Wang, Y. Xu, S. Tsogkas, X. Bai, S. Dickinson, and K. Siddiqi, "Deepflux for skeletons in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5287–5296.

[47] W. Ke, J. Chen, J. Jiao, G. Zhao, and Q. Ye, "Srn: Side-output residual network for object symmetry detection in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1068–1076.

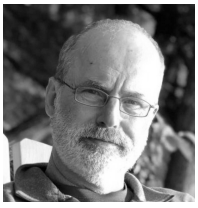
[48] X. Liu, P. Lyu, X. Bai, and M.-M. Cheng, "Fusing image and segmentation cues for skeleton extraction in the wild," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 1744–1748.

[49] J. H. Elder, T. D. Oleskiw, and I. Fruend, "The role of global cues in the perceptual grouping of natural shapes," *Journal of Vision*, vol. 18, no. 12, pp. 14–14, 11 2018.

Morteza Rezanejad received his bachelor's degree in computer engineering from the Sharif University of Technology, Tehran, Iran and his masters and his M.Sc. and Ph.D. degrees in computer science from McGill University, where he was a member of the shape analysis group in the Centre for Intelligent Machines. He held a position as a postdoctoral fellow in the Faculty of Arts and Science at the University of Toronto. His research interests are in computer vision, machine learning and shape analysis. The work presented in the current article was carried out while he was a doctoral student at McGill and a postdoc at the University of Toronto.

John Wilder was a research associate in the Department of Psychology at the University of Toronto while this work was being performed. He is now an assistant teaching professor at Northeastern University. Previously he worked as a postdoc at the University of Toronto in the Department of Psychology, and before that for the Department of Computer Science. John also did a postdoc at York University in the Centre for Vision Research. John completed his PhD in Psychology and an MS in Computer Science from Rutgers University. John earned a BA in Computers Science and Psychology at St. John's University in Minnesota.

Dirk B. Walther studied physics and computer science at the University of Leipzig. He received an M.Phil. degree in physics from the University of Cambridge, UK, in 1999 and a Ph.D. in Computation and Neural Systems from the California Institute of Technology in 2006. He was a postdoc at York University and a Beckman Postdoctoral Fellow at the University of Illinois at Urbana-Champaign, where he worked with Diane Beck and Fei-Fei Li on natural scene perception and on decoding natural scene categories from fMRI data. From 2010 until 2014, he was an Assistant Professor of Psychology at The Ohio State University. In 2014, he moved to the University of Toronto, where he is now Associate Professor of Psychology. He was a Visiting Professor at the Samsung Artificial Intelligence Center in Toronto from 2019 until 2020. His research focuses on the neural mechanisms that underlie the perception of complex real-world scenes both in human and computer vision. Recent research interests include the cross-modal perception of real-world environments as well as visual aesthetics. He is a senior member of the IEEE.



Allan D. Jepson received his B.Sc. in 1976 from the University of British Columbia and his Ph.D. in Applied Mathematics in 1980 from the California Institute of Technology (Caltech). He then moved to a postdoctoral position in the Mathematics Department at Stanford University. In 1982 he joined the faculty at the Department of Computer Science at the University of Toronto, becoming a full professor in 1991. Dr. Jepson was an Associate of the Canadian Institute of Advanced Research (CIFAR) for 1986 to 1989, for 2004 to 2009, and was a Scholar at CIFAR for 1989 to 1995. He served as Area Chair at the 2003 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), at the 2005 International Conference on Computer Vision (ICCV), and at the 2010 and 2018 European Conference on Computer Vision (ECCV). In 2016 he received the Lifetime Research Achievement Award from the Canadian Image Processing and Pattern Recognition Society (CIPPRS). He joined Samsung's AI Center in Toronto (SAIC-Toronto) in Sept. 2018 as Chief Scientist.

Sven Dickinson received the B.A.Sc. degree in Systems Design Engineering from the University of Waterloo, in 1983, and the M.S. and Ph.D. degrees in Computer Science from the University of Maryland, in 1988 and 1991, respectively. He is Professor and past Chair of the Department of Computer Science at the University of Toronto, and is also Vice President and Head of the new Samsung Toronto AI Research Center, which opened in May, 2018. Prior to that, he was a faculty member at Rutgers University where he held a joint appointment between the Department of Computer Science and the Rutgers Center for Cognitive Science (RuCCS). His research research interests revolve around the problem of shape perception in computer vision and, more recently, human vision. He has received the National Science Foundation CAREER award, the Government of Ontario Premiere's Research Excellence Award (PREA), and the Lifetime Research Achievement Award from the Canadian Image Processing and Pattern Recognition Society (CIPPRS). He was the Editor-in-Chief of the IEEE Transactions on Pattern Analysis and Machine Intelligence, from 2017-2021, currently serves on seven editorial boards, and is co-editor of the Morgan & Claypool Synthesis Lectures on Computer Vision. He is a Fellow of the International Association for Pattern Recognition (IAPR), and an IEEE Golden Core Member.

Kaleem Siddiqi received his BS degree from Lafayette College in 1988 and his MS and PhD degrees from Brown University in 1990 and 1995, respectively, all in the field of electrical engineering. He is currently a Professor at the School of Computer Science at McGill University where he holds an FRQS Dual Chair in Health and Artificial Intelligence. He is also a member of McGill's Centre for Intelligent Machines, an associate member of McGill's Department of Mathematics and Statistics, MILA - the Québec AI Institute, and the Goodman Cancer Centre. He presently serves as Field Chief Editor of the journal Frontiers in Computer Science. Before moving to McGill in 1998, he was a postdoctoral associate in the Department of Computer Science at Yale University (1996-1998) and held a position in the Department of Electrical Engineering at McGill University (1995-1996). More recently he was also a visiting professor and consultant at the Samsung AI Centre in Montréal (2021-2023). His research interests are in computer vision, robotics, image analysis, biological shape, neuroscience and medical imaging. He is a member of Phi Beta Kappa, Tau Beta Pi, and Eta Kappa Nu. He is a senior member of the IEEE.