Scene Categorization from Contours: Medial Axis Based Salience Measures

Morteza Rezanejad[1], Gabriel Downs[1], John Wilder[2], Dirk B. Walther[2],
Allan Jepson[2,3], Sven Dickinson[2,3], and Kaleem Siddiqi[1]

[1] McGill University, Montréal, QC, Canada

[2] University of Toronto, ON, Canada

[3] Samsung Toronto AI Research Center, ON, Canada*

Abstract

The computer vision community has witnessed recent advances in scene categorization from images, with the state of the art systems now achieving impressive recognition rates on challenging benchmarks. Such systems have been trained on photographs which include color, texture and shading cues. The geometry of shapes and surfaces, as conveyed by scene contours, is not explicitly considered for this task. Remarkably, humans can accurately recognize natural scenes from line drawings, which consist solely of contour-based shape cues. Here we report the first computer vision study on scene categorization of line drawings derived from popular databases including an artist scene database, MIT67 and Places365. Specifically, we use off-the-shelf pre-trained Convolutional Neural Networks (CNNs) to perform scene classification given only contour information as input, and find performance levels well above chance. We also show that medial-axis based contour salience methods can be used to select more informative subsets of contour pixels, and that the variation in CNN classification performance on various choices for these subsets is qualitatively similar to that observed in human performance. Moreover, when the salience measures are used to weight the contours, we find that these weights boost our CNN performance above that for unweighted contour input. That is, the medial axis based salience weights appear to add useful information that is not available when CNNs are trained to use contours alone.

1. Introduction

Both biological and artificial vision systems are confronted with a potentially highly complex assortment of visual features in real-world scenarios. The features need to be sorted and grouped appropriately in order to support high-level visual reasoning, including the recognition or categorization of objects or entire scenes. In fact, 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 whom are situated next to a body of water. A street scene might have roads with cars, cyclists and pedestrians as well as buildings along the edge. How might computer vision systems tackle this problem of organizing visual features to support scene categorization?

In human vision, perceptual organization is thought to be effected by a set of heuristic grouping rules originating from Gestalt psychology [13]. 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. Developed on an ad-hoc, heuristic basis originally, these rules have been validated empirically, even though their precise neural mechanisms remain elusive. Grouping cues, such as those based on symmetry, are thought to aid in high-level visual tasks such 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 [23], but the connection to the recognition of individual objects is not as straightforward as it may appear.

In computer vision, symmetry, proximity, good continuation, contour closure and other cues have been used for image segmentation, curve inference, object recognition, object manipulation, and other tasks [14, 2, 7, 17]. Instantiations of such organizational principles have found their way into many computer vision algorithms and have been the subject of regular workshops on perceptual organization in artificial vision systems. However, perceptually motivated

*Dr. Jepson and Dr. Dickinson contributed to this article in their personal capacity as Professors at the University of Toronto. The views expressed [or the conclusions reached] are their own and do not necessarily represent the views of Samsung Research America, Inc.

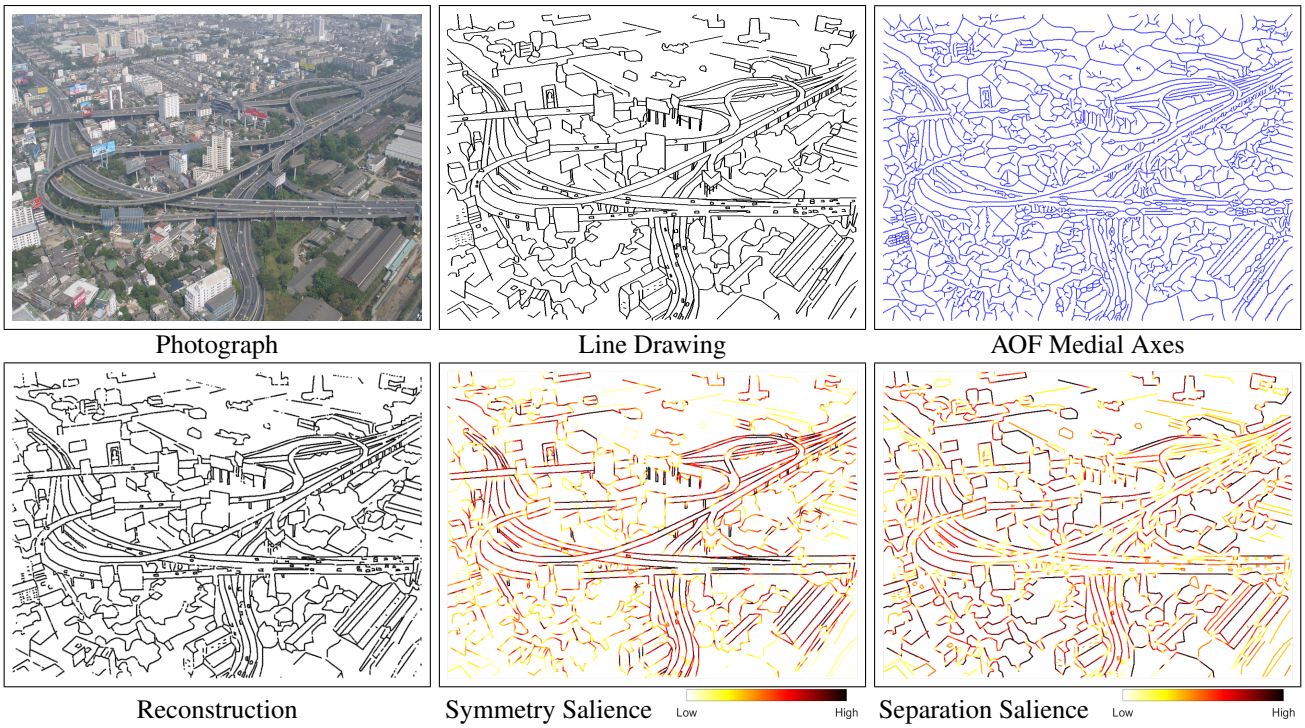

Figure 1: (Best viewed by zooming in on the PDF.) An illustration of our approach on an example from a database of line drawings by artists of photographs of natural scenes. The bottom left panel shows the reconstruction of the artist-generated line drawing from the AOF medial axes. To its right we present a hot colormap visualization of two of our medial axis based contour salience measures.

salience measures to facilitate scene categorization have received little attention thus far. 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 [18, 21, 11, 24]. 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 categories [20, 1, 8, 16]. Unfortunately, present CNN architectures do not explicitly allow for properties of object shape to be represented explicitly. Human observers, in contrast, recognize an object's shape as an inextricable aspect of its properties, along with its category or identity [12].

Comparisons between CNNs and human and monkey neurophysiology appear to indicate that CNNs replicate the entire visual hierarchy [9, 4]. 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 grouping 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 choose to use the medial axis transform (MAT) as a representation. We apply a robust algorithm for computing the medial axis to analyze line drawings of scenes of increasing complexity. The algorithm uses the average outward flux of the gradient of the Euclidean distance function through shrinking circular disks [5]. With its explicit representation of the regions between scene contours, the medial axis allows us to directly capture salience measures related to local contour separation and local contour symmetry. We introduce two novel measures of local symmetry using ratios of length functions derived from the medial axis radius along skeletal segments. 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. We describe methods of computing our perceptually motivated salience measures from line drawings of photographs of complex real-world scenes, covering databases of increasing complexity. Figure 1 presents an illustrative example of a photograph

from an artist scenes database, along with two of our medial axis based contour salience maps. Observe how the ribbon symmetry based measure highlights the boundaries of highways. Our experiments show that scene contours weighted by these measures can boost CNN-based scene categorization accuracy, despite the absence of colour, texture and shading cues. Our work indicates that measures of contour grouping, that are simply functions of the contours themselves, are beneficial for scene categorization by computers, yet that they are not automatically extracted by state-of-the-art CNN-based scene recognition systems. 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 for further study.

2. Average Outward Flux Based Medial Axes

In Blum's grassfire analogy the medial axis is associated with the quench points of a fire that is lit at the boundary of a field of grass [3]. In the present paper, that boundary is the set of scene contours, and the field of grass is the space between them. An equivalent notion of the medial axis is that of the locus of centres of maximal inscribed disks in the region between scene contours, along with the radii of these disks. The geometry and methods for computing the medial axis that we leverage are based on a notion of average outward flux, as discussed in further detail below. We apply the same algorithm to each distinct connected region between scene contours. These regions are obtained by morphological operations to decompose the original line drawing.

Definition 2.1 *Assume an n-dimensional open connected region Ω, with its boundary given by $\partial\Omega \in \mathbb{R}^n$ such that $\bar{\Omega} = \Omega \cup \partial\Omega$. An open disk $D \in \mathbb{R}^n$ is a maximal inscribed disk in $\bar{\Omega}$ if $D \subseteq \bar{\Omega}$ but for any open disk D' such that $D \subset D'$, the relationship $D' \subseteq \bar{\Omega}$ does not hold.*

Definition 2.2 *The Blum medial locus or skeleton, denoted by $Sk(\Omega)$, is the locus of centers of all maximal inscribed disks in $\partial\Omega$.*

Topologically, $Sk(\Omega)$ consists of a set of branches, about which the scene contours are locally mirror symmetric, that join at branch points to form the complete skeleton. A skeletal branch is a set of contiguous regular points from the skeleton that lie between a pair of junction points, a pair of end points or an end point and a junction point. At regular points the maximal inscribed disk touches the boundary at two distinct points. As shown by Dimitrov *et al.* [5] medial axis points can be analyzed by considering the behavior of the average outward flux (AOF) of the gradient of the Euclidean distance function through the boundary of the connected region. Let R be the region with boundary ∂R, and let $\mathbf{N}$ be the outward normal at each point on the boundary ∂R. The AOF is given by the limiting value of $\frac{\int_{\partial R}\langle \dot{\mathbf{q}}, \mathbf{N}\rangle ds}{\int_{\partial R} ds}$, as the region is shrunk. Here $\dot{\mathbf{q}} = \nabla \mathbf{D}$, with $\mathbf{D}$ the Euclidean distance function to the connected region's boundary, and the limiting behavior is shown to be different for each of three cases: regular points, branch points and end points. When the region considered is a shrinking disk, at regular points of the medial axis the AOF works out to be $-\frac{2}{\pi}\sin\theta$, where θ is the object angle, the acute angle that a vector from a skeletal point to the point where the inscribed disk touches the boundary on either side of the medial axis makes with the tangent to the medial axis. This quantity is negative because it is the inward flux that is positive. Furthermore, the limiting *AOF* value for all points not located on the medial axis is zero.

This provides a foundation for both computing the medial axis for scene contours and for mapping the computed medial axis back to them. First, given the Euclidean distance function from scene contours, one computes the limiting value of the AOF through a disk of shrinking radius and associates locations where this value is non-zero with medial axis points (Figure 1, top right). Then, given the AOF value at a regular medial axis point, and an estimate of the tangent to the medial axis at it, one rotates the tangent by $\pm\theta$ and then extends a vector out on either side by an amount given by the radius function, to reconstruct the boundary (Figure 1, bottom left). In our implementations we discretize these computations on a fine grid, along with a dense sampling of the boundary of the shrinking disk, to get high quality scene contour representations. Both the Euclidean distance function and the average outward flux computation are linear in the number of contour pixels, and thus can be implemented efficiently.

3. Medial Axis Based Contour Saliency

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 salience measures based on local contour separation and local symmetry. A measure to reflect local contour separation can be designed using the radius function along the medial axis, since this gives the distance to the two nearest scene contours on either side. Local parallelism between scene contours, or ribbon symmetry, can also be directly captured by examining the degree to which the radius function along the medial axis between them remains locally constant. Finally, if taper is to be allowed between contours, 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. We introduce novel measures to capture local separation, ribbon symmetry and taper, based on these ideas.

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

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

3.1. Separation Salience

We now introduce a salience 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 salience measure:

$$S_{Separation} = 1 - \Big(\int_{\alpha}^{\beta} \frac{1}{R(p)} dp \Big) / (\beta - \alpha). \quad (2)$$

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 exhibit further (local) separation are more salient by this measure.

3.2. Ribbon Symmetry Salience

Now consider the curve $\Psi = (x(p), y(p), R(p))$. Similar to Equation 1, the arc length of Ψ is computed as:

$$L_{\Psi} = \int_{\alpha}^{\beta} ||\frac{\partial \Psi}{\partial p}|| dp = \int_{\alpha}^{\beta} (x_p^2 + y_p^2 + R_p^2)^{\frac{1}{2}} dp. \quad (3)$$

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

$$S_{Ribbon} = \frac{L}{L_{\Psi}} = \frac{\int_{\alpha}^{\beta} (x_p^2 + y_p^2)^{\frac{1}{2}} dp}{\int_{\alpha}^{\beta} (x_p^2 + y_p^2 + R_p^2)^{\frac{1}{2}} dp}. \quad (4)$$

This quantity also falls in the interval $[0, 1]$ and is invariant to image scaling since the integral involves a ratio of unitless quantities. The measure is designed to increase as the scene contours on either side become more parallel, such as the two sides of a ribbon.

3.3. Taper Symmetry Salience

A notion that is closely related to that of ribbon symmetry is taper symmetry; 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 in 3D 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 symmetry salience measure:

$$S_{Taper} = \frac{L}{L_{\Psi'}} = \frac{\int_{\alpha}^{\beta} (x_p^2 + y_p^2)^{\frac{1}{2}} dp}{\int_{\alpha}^{\beta} (x_p^2 + y_p^2 + (RR_{pp})^2)^{\frac{1}{2}} dp}. \quad (5)$$

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.

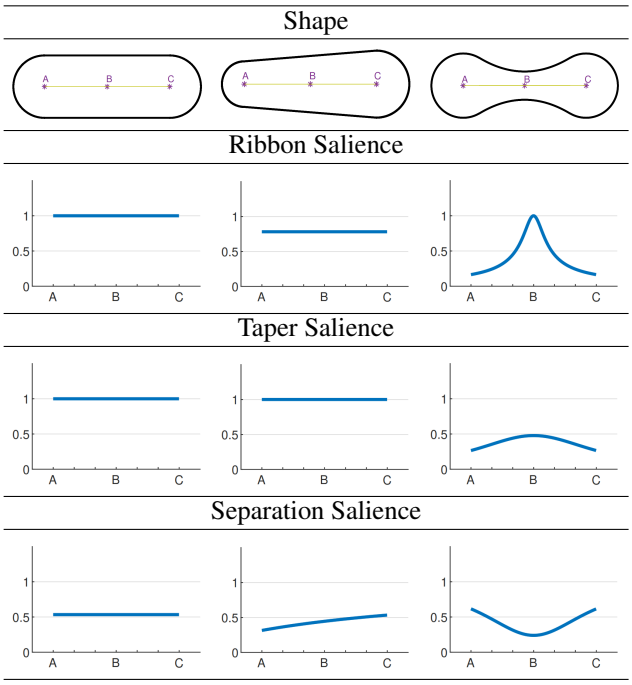

Figure 2: An illustration of ribbon symmetry salience, taper symmetry salience and contour separation salience for three different contour configurations. See text for a discussion. These measures are all invariant to 2D similarity transforms of the input contours

To gain an intuition behind these perceptually driven contour salience measures, we provide three illustrative examples in Fig. 2. The measures are not computed pointwise, but rather for a small interval $[\alpha, \beta]$ centered at each medial axis point (see Section 4.3 for details). When the contours are parallel, all three measures are constant along the medial axis (left column). The middle figure has high taper symmetry but lower ribbon symmetry, with contour separation salience increasing from left to right. Finally, for the dumbbell shape, all three measures vary (third column).

4. Experiments & Results

4.1. Artist Generated Line Drawings

Artist Scenes Database: Color 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. Line drawings of these photographs were generated by trained artists at the Lotus Hill Research Institute [22]. 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 paper, 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 79-80 exemplars from each of 6 categories: beaches, mountains, forests, highway scenes, city scenes and office scenes.

4.2. Machine Generated Line Drawings

MIT67/Places365 Given the limited number of scene categories in the Artist Scenes database, particularly for computer vision studies, we worked to extend our results to the two popular but much larger scene databases of photographs - MIT67 [15] (6700 images, 67 categories) and Places365 [24] (1.8 million images, 365 categories). Producing artist generated line drawings on databases of this size was not feasible, so instead we fine tuned the output of the Dollar edge detector [6], 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 then processed to obtain contour fragments of width 1 pixel. Each contour fragment was then spatially smoothed by convolution of the coordinates of points along it, using a Gaussian with $\sigma = 1$, to mitigate discretization artifacts. The same parameters were used to produce all the MIT67 and Places365 line drawings. Figure 3 presents a comparison of a resultant machine-generated and an artist-generated line drawing for an office scene from the Artist Scenes database. We have confirmed that on the artist's line drawing database 90% of the machine generated contour pixels are in common with the artist's line drawings. Figure 4 shows several typical machine generated line drawings from the MIT67 and Places365 databases, but weighted by our perceptual salience measures.

4.3. Computing Contour Salience

Computing contour salience for each line drawing required a number of steps. First, each connected region between scene contours was extracted. Second, we computed

Figure 3: (Best viewed by zooming in on the PDF.) A comparison between a machine-generated line drawing and one drawn by an artist, for an office scene from the Artist Scenes database.

an AOF map for each of these connected components, as explained in Section 2. For this we used a disk of radius 1 pixel, with 60 discrete sample points on it, to estimate the AOF integral. We used a threshold of $\tau = 0.25$ on the AOF map, which corresponds to an object angle $\theta \approx 23$ degrees, to extract skeletal points. A typical example appears in Figure 1 (top right). The resulting AOF skeleton was then partitioned into medial curves between branch points or between a branch point and an endpoint. We then computed a discrete version of each of the three salience measures in Section 3, within a interval $[\alpha, \beta]$ of length $2K+1$, centered at each medial axis point, with $K = 5$ pixels. Each scene contour point was then assigned the maximum of the two salience values at the closest points on the medial curves on either side of it, as illustrated in Figure 1 (bottom middle and bottom right).

4.4. Experiments on 50-50 Splits of Contour Scenes

Our first set of experiments is motivated by recent work that shows that human observers benefit from contour symmetry in scene recognition from contours [23]. Our goal is to examine whether a CNN-based system also benefits from such perceptually motivated cues. Accordingly, we created splits of the top 50% and the bottom 50% of the contour pixels in each image of the Artist Scenes and MIT67 data sets, using the three salience measures, ribbon symmetry, taper symmetry and local contour separation. An example of the original intact line drawing and each of the three sets of splits is shown in Figure 5, for the highway scene from the Artist Scenes dataset shown in Figure 1.

On the Artist Scenes dataset human observers were tasked with determining to which of six scene categories an exemplar belonged. The input was either the artist-generated line drawing or the top or the bottom half of a split by one of the salience 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. The results with these short image presentation durations, shown in Figure 6 (top), demonstrate that human performance is consistently better with the top (more salient) half of each split than the bottom one, for

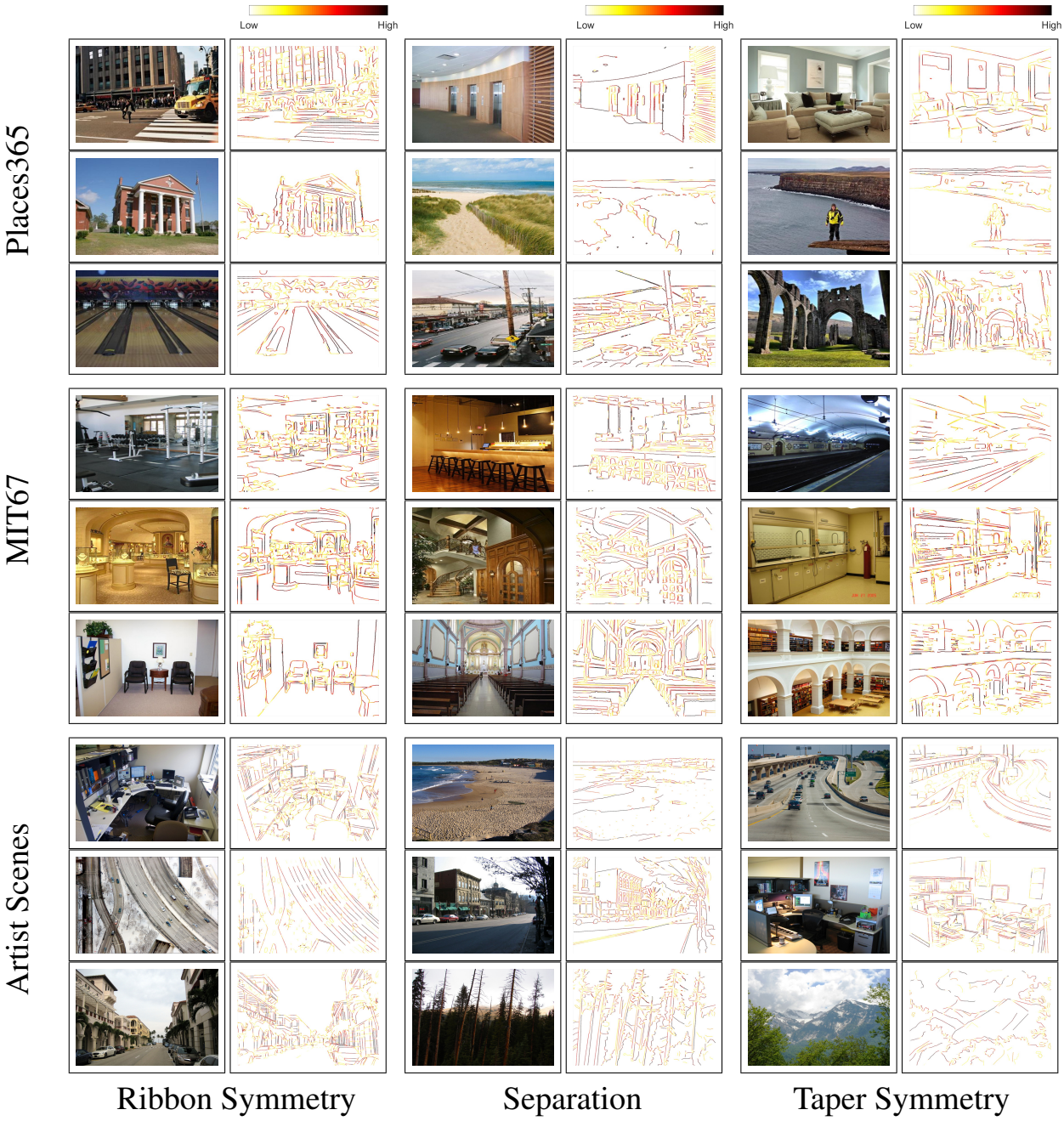

Figure 4: (Best viewed by zooming in on the PDF.) Examples of original photographs and the corresponding ribbon symmetry salience weighted, separation salience weighted and taper symmetry salience weighted scene contours, using a hot colormap to show increasing values. Whereas the Artist Scenes line drawings were produced by artists, the MIT67 and Places365 line drawings were machine-generated.

each salience measure. The human performance is slightly boosted for all conditions in the separation splits, for which a different subject pool was used.

Carrying out CNN-based recognition on the Artist Scenes 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 and MIT67 datasets,

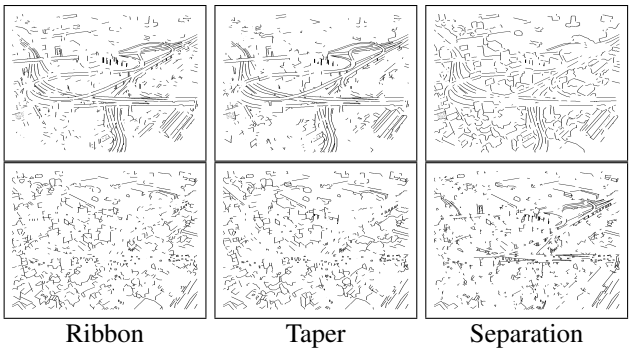

Figure 5: We consider the same highway scene as in Figure 1 (top left) and create splits of the artist generated line drawings, each of which contains 50% of the original pixels, based on ribbon symmetry (left column), taper symmetry (middle column) and local contour separation (third column) based salience measures. In each case the more salient half of the pixels is in the top row.

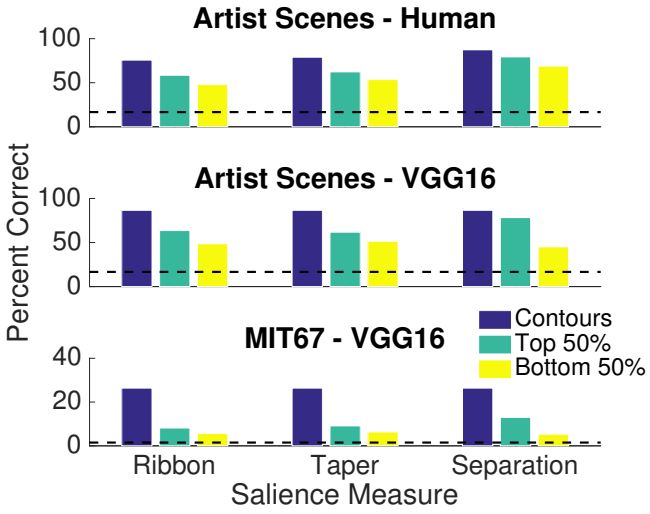

Figure 6: A comparison of human scene categorization performance (top row) with CNN performance (middle and bottom rows). As with the human observer data, CNNs perform better on the top 50% half of each split according to each salience measure, that the bottom 50% half. In each plot chance level performance (1/6 for Artist Scenes and 1/67 for MIT67) is shown with a dashed line.

we use the VGG16 convolutional layer network architecture [19] with weights pre-trained on ImageNet. The last three layers of the VGG16 network used for fine-tuning are replaced with a fully connected layer, a softmax layer and a classification 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

CNN	Human
Ribbon Sym vs Asymm $t(4) = 26.12$ $p = 1.3E{-}5$	Ribbon Sym vs Asymm $t(25) = 7.86$ $p = 3.2E{-}8$
Taper Sym vs Asym $t(4) = 12.39$ $p = 2.4E{-}4$	Taper Sym vs Asym $t(25) = 6.46$ $p = 9.2E{-}7$
Separation Far vs Near $t(4) = 100.64$ $p = 5.85E{-}8$	Separation Far vs Near $t(5) = 5.2$ $p = 3.0E{-}3$

Table 1: T-tests results for CNN and human categorization experiments.

as the prediction output. For the Places365 dataset, which contains 1.8 million images, we used Resnet50 [10] with its weights obtained by training on ImageNet, but rather than fine-tune the network, we used the final fully connected layer output as a feature vector input to an SVM classifier. For all experiments on the Artist Scenes we use 5-fold cross validation. Top-1 classification accuracy is given, as a mean over the 5 folds, in Figure 6 (middle). The CNN-based system mimics the trend we saw in human observers, namely that performance is consistently better for the top 50% of each of the three splits. We interpret this as evidence that all three Gestalt motivated salience 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 categorization results are shown in Figure 6 (bottom row). It is striking that even for this more challenging database, the CNN-based system still mimics the trend we saw in human observers, i.e., that performance is better on the top 50% than on the bottom 50% of each of the three splits and is well above chance. For both the CNN and human categorization experiments, we run t-tests (see Table 1) which show that for both the group differences are statistically significant.

4.5. Experiments with Salience 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 favor of our salience measures to support scene categorization. Can we exploit this bias to improve network performance when given the intact contours? To address this question, we carry out a second experiment where we explicitly encode salience measures for the CNN by feeding different features into the R, G, and B color channels of the pre-trained network. We do this by using, in addition to the contour image channel, additional channels with the same contours weighted by our proposed salience measures, each of which is in the interval $[0,1]$.

Channels	Artist	MIT67
	VGG16	VGG16
Photos	98.95	64.87
Contours	90.53	42.80
Contours, Ribbon	93.49	45.24
Contours, Taper	94.71	43.66
Contours, Separ.	93.91	43.89
Contours, Ribbon, Taper	95.02	45.36
Contours, Ribbon, Separ.	95.89	**48.61**
Contours, Taper, Separ.	**96.23**	47.18
Ribbon, Taper, Separ.	94.38	44.82

Table 2: Top 1 level performance in a 3-channel configuration, on Artist Scenes and MIT67, with fine-tuning. TOP ROW: Results of the traditional R,G,B input configuration. OTHER ROWS: Combinations of intact scene contours, and scene contours weighted by our salience measures.

Channels	Places365 (Res50)
Photos	33.04
Contours	8.02
Contours, Ribbon	9.18
Contours, Taper	11.73
Contours, Separ	10.53
Contours, Ribbon, Taper	12.05
Contours, Ribbon, Separ	**14.23**
Contours, Taper, Separ	11.77
Ribbon, Taper, Separ	12.64

Table 3: Top 1 performance in a 3-channel configuration on Places365, with an off-the-shelf pre-trained network and a linear SVM (see text). The top row shows the results of the traditional R,G,B input configuration, while the others show combinations of intact scene contours and scene contours weighted by our salience measures.

These contour salience images replace the standard three channel (R,G,B) inputs to the network. For all experiments, training is done on the feature maps generated by the new feature-coded images.

The results for the Artist Scenes dataset and for MIT67, are shown in Table 2. It is apparent that with these salience weighted contour channels added, there is a consistent boost to the results obtained by using contours alone. In all cases the best performance boost comes from a combination of contours, ribbon or taper symmetry salience, and separation salience. We believe this is because taper between local contours as a perceptual salience measure is conceptually very close to our ribbon salience measure. Local separation salience, on the other hand, provides a more distinct and complementary perceptual cue for grouping.

For MIT67 the performance of 64.87% on photographs is exactly consistent with that reported in [24]. Remarkably, two-thirds of this level of performance (42.8%) is obtained using *only* machine generated line drawings, and this goes up to three-fourths (48.6%) when using contours weighted by ribbon and separation salience. For MIT67 we have also compared (fine-tuned) Hybrid1365_VGG on photographs (**78.74% top-1**) versus photographs with contours, ribbon, and separation salience weighted contours overlayed (**80.45% top-1**).

Encouraged by the above results, we repeated the same experiment for the much more challenging Places365 dataset, but this time using just a pre-trained network and a linear SVM. For this dataset chance recognition performance would be at 1/365 or 0.27%. Our results are shown in Table 3. Once again we see a clear and consistent trend of a benefit using salience weighted contours as additional feature channels to the contours themselves, with the best performance gain coming from the addition of ribbon symmetry salience and separation salience.

5. Conclusion

We have reported the first study on CNN based recognition of complex natural scenes from line drawings derived from 3 databases of increasing complexity. To this end, we have demonstrated the clear benefit of using Gestalt motivated medial axis based salience measures, to weight scene contours according to their local ribbon and taper symmetry, and local contour separation. We hypothesize that making such contour salience weights explicit helps a deep network organize visual information to support categorization, in a manner which is not by default learned by these networks from the scene contour images alone. In our experiments, we used different CNN models to isolate the effect of these perceptually motivated scene contour grouping cues, and also the potential to perform scene categorization from contours alone, with color, shading and texture absent. In the artist's line drawings, 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.

The possibility to train a CNN model from scratch using our 1.8 million line drawings of Places365 now lies ahead. The feasibility of fully trained networks on drawings has been demonstrated by work on free-hand sketches [25], which, despite its superficial similarity with our work, follows a very different purpose. We plan on making our contour salience measure computation code, and our line drawing databases, publicly available.

Acknowledgments We are grateful to NSERC, Samsung, and Sony for research support.

References

[1] Shuang Bai. Growing random forest on deep convolutional neural networks for scene categorization. *Expert Systems with Applications*, 71:279–287, 2017. 2

[2] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological Review*, 94(2):115, 1987. 1

[3] Harry Blum. Biological shape and visual science (part i). *J. Theor. Biol.*, 38(2):205–287, Feb. 1973. 3

[4] Charles F Cadieu, Ha Hong, Daniel LK Yamins, Nicolas Pinto, Diego Ardila, Ethan A Solomon, Najib J Majaj, and James J DiCarlo. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Computational Biology*, 10(12):e1003963, 2014. 2

[5] Pavel Dimitrov, James N Damon, and Kaleem Siddiqi. Flux invariants for shape. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1, pages I–835. IEEE, 2003. 2, 3

[6] Piotr Dollár and C. Lawrence Zitnick. Structured forests for fast edge detection. In *ICCV*, 2013. 5

[7] James H Elder and Steven W Zucker. Computing contour closure. In *European Conference on Computer Vision*, pages 399–412. Springer, 1996. 1

[8] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014. 2

[9] Umut Güçlü and Marcel A. J. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015. 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 7

[11] Shin Hoo-Chang, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 35(5):1285, 2016. 2

[12] Philip J Kellman and Thomas F Shipley. A theory of visual interpolation in object perception. *Cognitive Psychology*, 23(2):141–221, 1991. 2

[13] Kurt Koffka. Perception: an introduction to the gestalt-theorie. *Psychological Bulletin*, 19(10):531, 1922. 1

[14] David Marr and Herbert Keith Nishihara. Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 200(1140):269–294, 1978. 1

[15] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 413–420. IEEE, 2009. 5

[16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015. 2

[17] Sudeep Sarkar and Kim L Boyer. Perceptual organization in computer vision: status, challenges, and potential. *Computer Vision and Image Understanding*, 76(1):1–5, 1999. 1

[18] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 806–813, 2014. 2

[19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 7

[20] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun RGB-D: A RGB-D scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 567–576, 2015. 2

[21] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[22] Dirk B Walther, Barry Chai, Eamon Caddigan, Diane M Beck, and Li Fei-Fei. Simple line drawings suffice for functional mri decoding of natural scene categories. *Proceedings of the National Academy of Sciences*, 108(23):9661–9666, 2011. 5

[23] John Wilder, Morteza Rezanejad, Sven Dickinson, Kaleem Siddiqi, Allan Jepson, and Dirk B. Walther. Local contour symmetry facilitates scene categorization. *Cognition*, 182:307 – 317, 2019. 1, 5

[24] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018. 2, 5, 8

[25] Changqing Zou, Qian Yu, Ruofei Du, Haoran Mo, Y SONG, Tao Xiang, Chengying Gao, Baoquan Chen, Hao Zhang, et al. Sketchyscene: Richly-annotated scene sketches. European Conference on Computer Vision, 2018. 8