Gestalt-based Contour Weights Improve Scene Categorization by CNNs

Morteza Rezanejad[1], Gabriel Downs[1], John Wilder[2,3], Dirk B. Walther[3], Allan Jepson[2,4], Sven Dickinson[2,4], Kaleem Siddiqi[1]

[1]School of Computer Science & Centre for Intelligent Machines, McGill University, Montreal, PQ, Canada

[2]Department of Computer Science, University of Toronto, Toronto, ON, Canada

[3]Department of Psychology, University of Toronto, Toronto, ON, Canada

[4]Samsung Toronto AI Research Center, Toronto, ON, Canada

Abstract

Humans can accurately recognize natural scenes from line drawings, consisting solely of contour-based shape cues. Deep learning strategies for this complex task, however, have thus far been applied directly to photographs, exploiting all the cues available in colour images at the pixel level. Here we report the results of fine tuning off-the-shelf pre-trained Convolutional Neural Networks (CNNs) to perform scene classification given only contour information as input. To do so we exploit the Iverson-Zucker logical/linear framework to obtain line drawings from popular scene categorization databases, including an artist's scene database and MIT67. We demonstrate a high level of performance despite the absence of colour, texture and shading information. We also show that the inclusion of medial-axis based contour salience weights leads to a further boost, adding useful information that does not appear to be exploited when CNNs are trained to use contours alone.

Keywords: Scene Categorization, Logical/Linear Line Drawings, Perceptual Grouping, Medial Axis, Contour Salience, Contour Symmetry, Contour Separation

Introduction

In vision science perceptual organization is thought to be effected by a set of heuristic grouping rules originating from Gestalt psychology (Koffka, 1922). 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 (Wilder et al., 2019), but the connection to the recognition of individual objects is not as straightforward as it may appear.

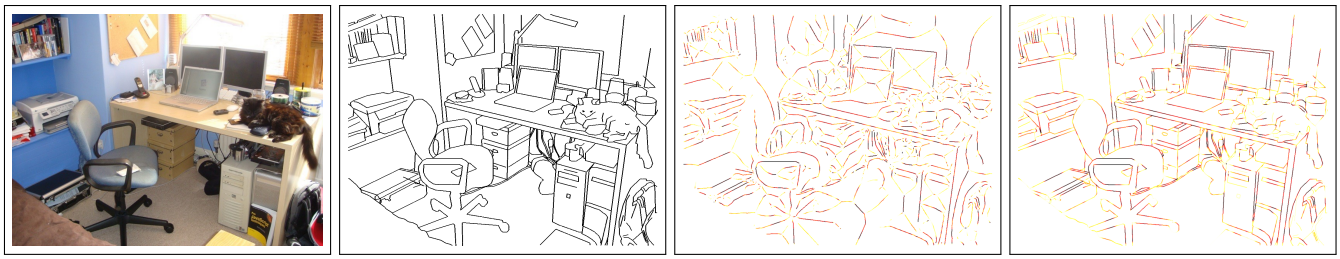
In computer vision, symmetry, proximity, good continuation, contour closure and other cues have been used for image segmentation, curve inference, object recognition, object manipulation, and other tasks (Marr & Nishihara, 1978; Biederman, 1987; Elder & Zucker, 1996; Sarkar & Boyer, 1999).

However, perceptually motivated 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 (Sharif Razavian, Azizpour, Sullivan, & Carlsson, 2014; Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016; Zhou, Lapedriza, Khosla, Oliva, & Torralba, 2018). 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 (Song, Lichtenberg, & Xiao, 2015; Bai, 2017; Girshick, Donahue, Darrell, & Malik, 2014; Ren, He, Girshick, & Sun, 2015). Unfortunately, present CNN architectures do not 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 (Kellman & Shipley, 1991).

Comparisons between CNNs and human and monkey neurophysiology appear to indicate that CNNs replicate the entire visual hierarchy (Kriegeskorte, 2015; Kar, Kubilius, Schmidt, Issa, & DiCarlo, 2019; Güçlü & van Gerven, 2015; Cadieu et al., 2014). Does this mean that the problem of perceptual organization is now irrelevant for machine vision? We argue that this is not the case, and show that CNN-based scene categorization systems, just like human observers, can benefit from explicitly computed contour measures derived from Gestalt grouping cues. To do so we use an average outward flux formulation to compute the medial axis (Dimitrov, Damon, & Siddiqi, 2003) and then use it to directly capture salience measures related to local contour separation and local contour symmetry. 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.

Medial Axis Based Contour Saliency

Motivated by the considerations above, we recently introduced novel measures to capture local separation, ribbon symmetry and taper from line drawings of natural scenes (Rezanejad et al., in press), which we now review. Owing to the continuous mapping between the medial axis and scene contours, the scores obtained using these measures can then be mapped to scene contours. We let p be a parameter that runs along a



Scene | Line drawing | Salience score on the medial axis | Salience score on the line drawing

Figure 1: (Best viewed by zooming in on the PDF.) An illustration of our approach on an office scene example from the Artist Scenes Database. On the right we present a hot colormap visualization of our salience measures on the medial axis and the line drawing, respectively. Here black represents high salience, red intermediate salience, and yellow low salience.

medial axis segment, $(x(p), y(p))$ be the coordinates of points along that segment, and $R(p)$ be the medial axis radius at each point. We consider the interval $p \in [\alpha, \beta]$ for a particular medial segment.

Separation Salience

With $R(p) > 1$ in pixel units (we assume that two distinct scene contours do not touch) we use the following contour separation based salience measure:

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

This quantity falls in the interval $[0,1]$, increasing with greater spatial separation between the two contours. Scene contours that exhibit further (local) separation are more salient by this measure.

Ribbon Symmetry Salience

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:

$$S_{Ribbon} = \frac{\int_{\alpha}^{\beta} (x_p^2 + y_p^2)^{\frac{1}{2}} dp}{\int_{\alpha}^{\beta} (x_p^2 + y_p^2 + R_p^2)^{\frac{1}{2}} dp}. \qquad (2)$$

This measure also falls in the interval $[0,1]$ and it increases as the scene contours on either side become more parallel, such as the two sides of a ribbon.

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 projection. To capture this notion of symmetry we use the following taper symmetry salience measure:

$$S_{Taper} = \frac{\int_{\alpha}^{\beta} (x_p^2 + y_p^2)^{\frac{1}{2}} dp}{\int_{\alpha}^{\beta} (x_p^2 + y_p^2 + (RR_{pp})^2)^{\frac{1}{2}} dp}. \qquad (3)$$

This quantity also falls in the interval $[0,1]$ and it increases as the scene contours on either side become more taper symmetric, as in the sides of a railway track.

Experiments & Results

Artist Scenes Database

Color photographs of six categories of natural scenes were downloaded from the internet, and those rated as the best exemplars of their respective categories by workers on Amazon Mechanical Turk were selected ((Torralbo et al., 2013)). Line drawings of these photographs were generated by trained artists at the Lotus Hill Research Institute (Walther, Chai, Caddigan, Beck, & Fei-Fei, 2011). The resulting database had 475 line drawings across 6 categories: beaches, mountains, forests, highway scenes, city scenes and office scenes.

Machine Generated Logical/Linear Line Drawings

Given the limited number of scene categories in the Artist Scenes database, we worked to extend our analysis on a much larger scene database of photographs - MIT67 (Quattoni & Torralba, 2009) (6700 images, 67 categories). To produce line drawings from this much larger database we modified the output of the logical/linear edge detector (Iverson & Zucker, 1995), using their publicly available open source implementation. This approach is devised to recover image curves while preserving singularities and junctions. We briefly review the three kinds of image curves modeled in (Iverson & Zucker, 1995).

Consider an image $I : \mathbb{R}^2 \to \mathbb{R}^+$, with $P = [\alpha, \beta]$ and let $\mathbf{C} : p \in P \to \mathbb{R}^2$ represent a smooth curve parameterized by arc length. The normal cross section $\mathbf{N}_p(t)$ at the curve point $\mathbf{C}(p)$ is given by:

$$\mathbf{N}_p(t) = I(\mathbf{C}(p) + t\mathbf{N}(p))), \quad p \in P, \quad t \in \mathbb{R}. \qquad (4)$$

Using local structural conditions in the directions tangential and normal to the curve, the following three image curve categories are suggested in (Iverson & Zucker, 1995):

1. $\mathbf{C}$ is an *Edge* iff $\mathbf{C}$ is an image curve such that the following condition holds for all $p \in P$:

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

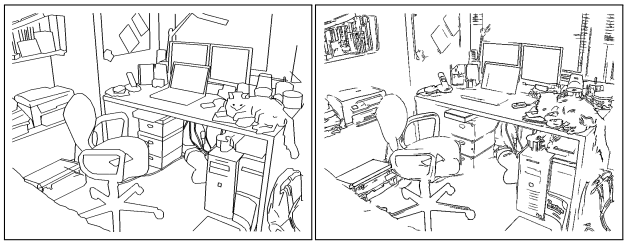
2. $\mathbf{C}$ is a *Positive Constrast Line* iff $\mathbf{C}$ is an image curve such that the following condition holds for all $p \in P$:

$$\lim_{t \to 0^-} \mathbf{N}'_p(t) > 0 \text{ and } \lim_{t \to 0^+} \mathbf{N}'_p(t) < 0$$

3. $\mathbf{C}$ is a *Negative Constrast Line* iff $\mathbf{C}$ is an image curve such that the following condition holds for all $p \in P$:

$$\lim_{t \to 0^-} \mathbf{N'}_p(t) < 0 \text{ and } \lim_{t \to 0^+} \mathbf{N'}_p(t) > 0$$

In (Iverson & Zucker, 1995) operators are designed to respond when any of the above conditions are met locally in an image, and if so, either an edge, or a line is reported. In our experiments we focused on the case of edge points; from the output edge map and its associated edge strength and edge directions, we produced a binarized version. Each binarized edge map was processed and traced to obtain contour fragments having a width of 1 pixel. Figure 2 presents a comparison of an artist-generated line drawing for an office scene from the Artist Scenes database, along with the logical/linear (machine generated) version.



Artist Line Drawing Machine Generated Line Drawing

Figure 2: (Best viewed by zooming in on the PDF.) An artist's line drawing of an office scene, and the machine generated version, obtained using logical/linear operators (Iverson & Zucker, 1995).

We have confirmed that on the artist's line drawing database 82% of the machine generated contour pixels are in common with the artist's line drawings.

Scene Categorization with Salience Weighted Contours

We report the results of scene categorization using both contours and contours weighted with our perceptual salience measures. We accomplish this by feeding different features in the 3 channels (normally used for red, green and blue) of a pre-trained network, as illustrated in Figure 3. We have used VGG16 (pre-trained on Imagenet) and VGG16-H (pre-trained on both Imagenet and Places365 (Zhou et al., 2018)). In all the experiment,s the last two fully-connected layers of the pre-trained networks were fine-tuned using our feature-coded inputs, i.e., training was done on the feature maps provided by them.

The results for the Artist Scenes dataset and for MIT67, are shown in Table 1. 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 biggest performance boost comes from a combination of contours, ribbon or taper symmetry salience, and separation salience. This is likely because taper salience is conceptually very close to ribbon salience, while local separation salience provides a more distinct and complementary perceptual cue for grouping.

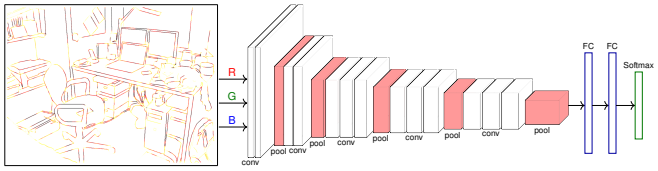


Figure 3: (Best viewed by zooming in on the PDF.) A schematic view of the VGG16 architecture with salience weighted contours used as the 3 input channels (see Table 1 for the specific sets of input channels.

Channels	Artist		MIT67	
	VGG16	VGG16-H	VGG16	VGG16-H
Photos	98.20	99.62	64.87	79.49
CCC	91.23	92.50	46.92	60.73
CCR	93.46	94.16	48.55	61.10
CCT	93.10	95.06	49.84	63.32
CCS	94.63	96.56	49.61	62.54
CRT	94.85	96.61	51.32	62.96
CRS	95.42	**98.40**	53.21	64.25
CTS	**96.82**	97.93	**54.17**	**65.79**
RTS	95.74	95.96	52.52	63.48

Table 1: Top 1 level performance in a 3-channel configuration, on the Artist Scenes 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 salience measures, where each letter stands for a specific input channel. **C** = contours, **R** = ribbon symmetry, **T** = taper symmetry and **S** = separation.

For MIT67 the performance of 79.49% on photographs is consistent with that reported in (Zhou et al., 2018). Remarkably, 75% of this level of performance (a level of 60.73%) is obtained using *only* logical/linear line drawings. The overall performance goes up to 65.79% (or 82.8% of the performance on photographs) when using contours weighted by ribbon and separation salience. For MIT67, we have also compared the performance (fine-tuned) Hybrid1365_VGG on photographs (**79.49% top-1**) with photographs with contours, ribbon, and separation salience weighted contours overlayed (**82.05% top-1**). Thus, perceptually weighted contour features can boost overall performance as well.

Conclusion

Our experiments show that scene contours weighted by perceptually motivated contour salience 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, which are simply functions of the contours themselves, are beneficial for scene catego-

rization by computers, leading to recognition performance that is over 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. Adding shape information computed on the medial axis outside of the CNNs improves scene categorization above the current state of the art.

Acknowledgments

We are grateful to NSERC, Samsung, and Sony for funding that has supported this research.

References

Bai, S. (2017). Growing random forest on deep convolutional neural networks for scene categorization. *Expert Systems with Applications*, *71*, 279–287.

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

Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., ... DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Computational Biology*, *10*(12), e1003963.

Dimitrov, P., Damon, J. N., & Siddiqi, K. (2003). Flux invariants for shape. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Vol. 1, pp. I–835).

Elder, J. H., & Zucker, S. W. (1996). Computing contour closure. In *European Conference on Computer Vision* (pp. 399–412).

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 580–587).

Güçlü, U., & van Gerven, M. A. J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

Iverson, L. A., & Zucker, S. W. (1995, Oct). Logical/linear operators for image curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *17*(10), 982-996. doi: 10.1109/34.464562

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, *22*(6), 974-983. doi: 10.1038/s41593-019-0392-5

Kellman, P. J., & Shipley, T. F. (1991). A theory of visual interpolation in object perception. *Cognitive Psychology*, *23*(2), 141–221.

Koffka, K. (1922). Perception: an introduction to the gestalt-theorie. *Psychological Bulletin*, *19*(10), 531.

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*(1), 417-446.

Marr, D., & Nishihara, H. K. (1978). 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.

Quattoni, A., & Torralba, A. (2009). Recognizing indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 413–420).

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems* (pp. 91–99).

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

Sarkar, S., & Boyer, K. L. (1999). Perceptual organization in computer vision: status, challenges, and potential. *Computer Vision and Image Understanding*, *76*(1), 1–5.

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

Song, S., Lichtenberg, S. P., & Xiao, J. (2015). Sun RGB-D: A RGB-D scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 567–576).

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016, June). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Torralbo, A., Walther, D. B., Chai, B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2013, 03). Good exemplars of natural scene categories elicit clearer patterns than bad exemplars but not greater bold activity. *PLOS ONE*, *8*(3), 1-12.

Walther, D. B., Chai, B., Caddigan, E., Beck, D. M., & Fei-Fei, L. (2011). Simple line drawings suffice for functional mri decoding of natural scene categories. *Proceedings of the National Academy of Sciences*, *108*(23), 9661–9666.

Wilder, J., Rezanejad, M., Dickinson, S., Siddiqi, K., Jepson, A., & Walther, D. B. (2019). Local contour symmetry facilitates scene categorization. *Cognition*, *182*, 307 - 317. doi: https://doi.org/10.1016/j.cognition.2018.09.014

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2018). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(6), 1452–1464.