
Proceedings Paper:
Kyle-Davidson, Cameron, Bors, Adrian Gheorghe orcid.org/0000-0001-7838-0021 and Evans, Karla orcid.org/0000-0002-8440-1711 (2022) Predicting Human Perception of Scene Complexity. In: IEEE International Conference on Image Processing (ICIP). IEEE , Bordeaux, France .

https://doi.org/10.1109/ICIP46576.2022.9897953

PREDICTING HUMAN PERCEPTION OF SCENE COMPLEXITY

Cameron Kyle-Davidson[⋆] *Adrian G. Bors*[⋆] *Karla K. Evans*[†]

[⋆]Dept. of Computer Science,[†]Dept. of Psychology, University of York, York, UK

ABSTRACT

It is apparent that humans are intrinsically capable of determining the degree of complexity present in an image; but it is unclear which regions in that image lead humans towards evaluating an image as complex or simple. Here, we develop a novel deep learning model for predicting human perception of the complexity of natural scene images in order to address these problems. For a given image, our approach, ComplexityNet, can generate both single-score complexity ratings and two-dimensional per-pixel complexity maps. These complexity maps indicate the regions of scenes that humans find to be complex, or simple. Drawing on work in the cognitive sciences we integrate metrics for scene clutter and scene symmetry, and conclude that the proposed metrics do indeed boost neural network performance when predicting complexity.

***Index Terms*—** deep learning, complexity perception, image analysis, cognitive science, human vision

1. INTRODUCTION

Humans are capable of rapidly evaluating the visual complexity of an image; it is immediately and obviously apparent that a painting of a scene is more complex than a blank canvas. Beyond this simplistic and stark contrast, there is significant evidence that human beings are capable of evaluating the complexity of images that vary from simple textures, to complex images of objects. Research into human complexity perception has both theoretical and practical aspects. From the theoretical side, determining which elements contribute to the level of complexity perceived in an image reveal clues as to how the human visual system operates and automatically evaluates stimuli. Practically, applications of studying complexity range from marketing to healthcare. An advert may need to be complex enough to hold attention and inform, but not so complex that it cannot be understood; the same rationale applies to educational materials. Psychological experiments may require stimuli of identical complexity to exclude a confounding factor, and how complex a person finds an image could even be used to track cognitive decline in visual processing disorders.

The study of perceptual complexity is not new; research goes back decades and spans the fields of both cognitive psychology and computer science. The theory of complexity perception originally began in the early 20th century as part of a study on the aesthetics of images [1]; suggesting that complexity is related to the number of distinct elements in an image. This was later redefined as the intricacy present in a line-drawn representation of an image [2], followed by the level of difficulty inherent in verbally describing a texture [3]. These results are indicative that humans can perceive the complexity of images, but these images are simplistic compared to the natural scenes in which we are constantly immersed. The first study to evaluate scenes in particular found that both the level of clutter present in a scene, and the symmetry of the scene play a role in complexity perception [4].

Advances in computing power have lead to definitions of visual complexity based on the information sciences; the Shannon entropy of an image is often used to measure complexity [5, 6]; more complex images are assumed to have a greater degree of disorganisation, and hence, greater entropy compared to simple images. This, and a related measure, the Kolmogorov complexity [7] (often approximated through compression algorithms) are viewed as measures of the clutter present in the image [8, 9] (clutter having been found to relate to overall complexity). However, informational metrics suffer from being somewhat divorced from human perception; an image of random noise has both high entropy and a high Kolmogorov complexity, yet is meaningless to a person. Recent approaches to complexity prediction either focus on combinations of several different metrics [10, 11, 12], or in employing neural network models to predict complexity scores [13]. One such study found that neural networks appear to automatically reveal the complex regions of images in an unsupervised fashion [13]; and this "unsupervised activation energy" (UAE) metric is capable of predicting the complexity of images. That is, the neurons of neural networks trained for classification automatically activate in the presense of complex features.

Up to this point, methods that predict complexity focus on generating a single score for a given image. These methods do not reveal the regions of images that lead humans to rate an image as more complex, or more simple. This is primarily due to the lack of ground-truth data. As such, the UAE metric could not be directly compared with human data to determine whether neural networks and human beings find the same parts of images complex. However, the community is now beginning to treat "perceptual image characteristics"

such as image memorability, as more than a single rating; instead investigating how said properties vary across an image [14, 15]. This has recently culminated in the development of the "VISC-C" dataset [16], which contains a variety of scene images, complexity scores, and complexity maps. However, unlike image memorability, there has not yet been development of any neural architectures capable of predicting complexity maps.

We introduce a deep learning network for predicting complexity for scene images. Said network can produce both complexity maps that highlight the image regions a human may find complex, or simplistic, and can also generate complexity scores for these scenes that can be directly compared with human ratings. Given that prior work has found that clutter and symmetry relate to complexity perception, we introduce an optional module that can feed this information into the network. We set a baseline for human complexity map prediction based on human data, and evaluate the relationship between the prior state-of-the-art method of complexity prediction, the unsupervised activation energy; comparing both the final score produced by the method, and the intermediate "UAE maps" that highlight regions that may be complex in the scene.

2. COMPLEXITYNET: PREDICTING PERCEIVED VISUAL COMPLEXITY

In the following we describe the deep learning network used to model complexity of scene images. Given evidence for clutter and symmetry and its relation to complexity perception [4], we experiment with including said measures as optional inputs for our proposed approach. Clutter can be loosely defined as the number of visually distinct regions present in an image; to capture this we employ a region-adjacency graph cut algorithm [17] to divide input scene images into a number of "perceptually distinct" regions. The normalised cut of graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ into regions A, B is

$$Ncut(A,B) = \frac{Cut(A,B)}{Assoc(A,V)} + \frac{Cut(A,B)}{Assoc(B,V)} \quad (1)$$

where $Cut(A,B)$ computes the sum of edge weights removed, and $Assoc(A,V)$ is the sum of edge weights from A to all vertices in the region-adjacency graph. This results in a segmentation into distinct regions based on their colour similarity. For symmetry extraction we compute the local symmetry of various patches throughout the image. The amount of perceptually redundant information increases with the presence of more local symmetries. This indicates a lower image complexity. We compute the symmetry of image patches as follows: Given patch $N_{ij}^{h\times w\times c}$, at location (i,j), bisect the patch vertically, resulting in $(A^{h\times \frac{w}{2}\times c}, B^{h\times \frac{w}{2}\times c})$, where $A_{ij} = N_{i,0<j<\frac{w}{2}}$ and $B_{ij} = N_{i,\frac{w}{2}<j<w}$, defining $F_h(A)$ as the horizontal flip of A, the horizontal symmetry of the patch is

$$sym_h(N) = \sqrt{(f_h(A) - B)^2} \quad (2)$$

Hence, $sym(N) = \frac{H_n^{sym} + V_n^{sym}}{2}$, and the overall symmetry of map of image I is

$$sym(I) = \sum_{k=0}^{K} sym(N_k^{h\times w\times c}) \quad (3)$$

where K is the set of patches extracted. This generates a symmetry map; with asymmetric regions of the image being assigned a lower value than symmetric image regions.

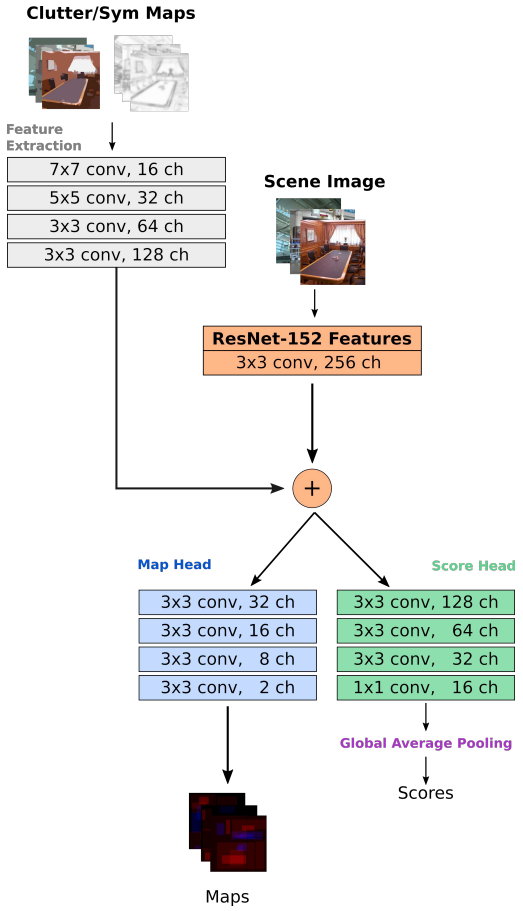

Fig. 1. Proposed dual-headed complexity map and score prediction network.

Inspired by recent work in neural network-based methods for predicting two-dimensional memorability maps [14, 18], we propose a deep learning model for predicting two-dimensional (i.e per-pixel) complexity maps. We term this model 'ComplexityNet'. Given that it's advantageous to have both complexity maps and complexity scores (maps provide the detail, scores a summary) our network includes two prediction heads that are optimised jointly; one that predicts map,

and one scores for the input images. We optionally include a module that can integrate features learnt from generated clutter and symmetry maps, given their hypothesised relation to complexity perception. The architecture of our proposed approach can be seen in Figure 1. We select a pretrained ResNet-152 [19, 20] object-detection backbone for our network under the hypothesis that the semantic features extracted by such a network are relevant for complexity perception. We truncate the backbone before the final classification layer, with extracted feature tensor $R^{H\times W\times C} \in \mathbb{R}$ where H, W, C are the height, width, and channels of the feature tensor; in this case, $7\times7\times2048$. The input to the network is a 3D tensor with seven channels, $\mathbf{X}^{224\times224\times7} \in \mathbb{R}$. Of these seven channels, 3 correspond to the RGB channels of the input scene image, 3 correspond to a *clutter map*, and 1 corresponds to a *symmetry* map.

The part of the input representing the scene image is passed through the ResNet feature extractor. However, it would be nonsensical to attempt to use this to extract features from clutter and symmetry maps; for this we design our own four-layer feature extractor to extract features from the clutter and symmetry map; a four channel tensor subset of the input. Learnt features from the clutter and symmetry maps are included in a "shared feature core" that concatenates these features with the ResNet-based semantic features. These features are then passed to the map/score prediction head. We use downsampling convolutions in the clutter/symmetry feature extractor for dimensionality reduction, avoiding pooling operations with no learnable features that may reduce relevant features accidentally given the small dataset size. We do the same for the score prediction head, culminating in a single average pooling operation to condense features maps to a singular complexity score. Throughout the network we keep the number of feature maps of each layer limited; not exceeding 256 in any one layer; again due to the small dataset size. The output of the network is a 1-dimensional complexity score and a 2-dimensional complexity map that gives a 2-channel per-pixel score for each pixel in the input image. The per-pixel score captures either the level of complexity, or the level of simplicity, that should be assigned to that pixel.

2.1. Loss Function

During the training of ComplexityNet we need to optimise both complexity scores, and complexity maps, framed as a regression problem from predicted ratings/maps to ground-truth. We compute the loss of the complexity maps over the simple and complex channels compared to the ground-truth human maps, and combine this with the loss of the score into the following loss function:

$$L = \sum_{i=1}^{C}(\mathbf{M}_i - \hat{\mathbf{M}}_i)^2 + \sum_{i=1}^{S}(S_i - \hat{S}_i)^2 \qquad (4)$$

where C represents the set of complexity maps, with M_i and $\hat{M}_i$ representing the ground truth and predicted map of that set respectively, and S the set of complexity ratings S_i and $\hat{S}_i$ the ground truth and predicted complexity ratings.

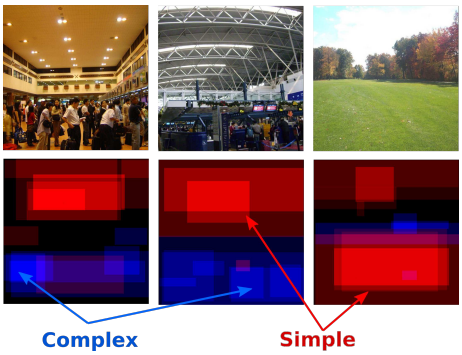

Fig. 2. Sample images and complexity maps from the VISC-C dataset. Red regions indicate areas of scenes humans find simple, and blue regions indicate complex regions.

3. EXPERIMENTAL RESULTS

3.1. Dataset

We use the publicly available [16] Vischema-Complexity (VISC-C) dataset, a collection of 800 700×700 scene images, with 800 associated 'complexity maps' that show the complex and simple regions of those scenes. Each scene also comes with an aggregate 'complexity score' giving a single numeric value for the visual complexity of that scene. The data was gathered from 40 human observers, with each individual scene being viewed by 10 participants in total. The image-set is divided into eight scene classes (kitchen, living room, conference room, airport terminal, work/home, public entertainment, populated and isolated outdoor scenes) of 100 images each, with each class corresponding to a commonly encountered real-world scene category.

3.2. Implementation and Training details

We implement our proposed approach using the PyTorch Machine Learning Library [21], in Python. We train the network using the RMSProp optimizer with a learning rate of 0.0001 for 200 epochs. We employ n-fold cross-validation during training, selecting 12.5% of the data as a test set and training on the rest. Given the small size of the dataset, during training we employ random data augmentation with a probability of 0.5. During augmentation, both the input image and all related input maps (complexity, clutter, symmetry) are flipped around the vertical axis. The clutter and symmetry input can be disabled or enabled as required; the network is fully capable of being trained on images alone. The network takes approximately 20 minutes to train on a single NVidia V100 GPU.

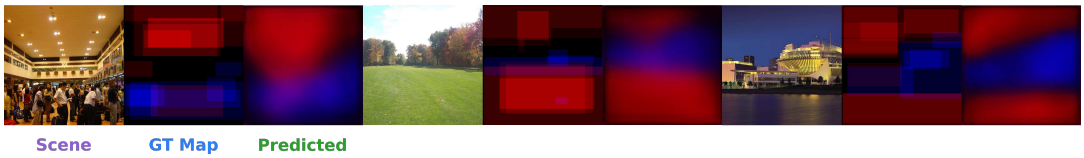

Fig. 3. Sample results images. GT Map refers to the ground-truth human data for that scene image, while Predicted shows results from our best performing network. The network appears to have learned the regions that humans find both simple, and complex, in natural scene images. Red regions are predicted simple, blue regions, complex.

Approach	Score	Complex Regions	Simple Regions
UAE [13]	0.397	0.36	-
CNet	0.693	0.51	0.37
CNet-C	**0.729**	0.5	0.38
CNet-S	0.719	0.51	0.38
CNet-CS	0.716	**0.51**	**0.39**

Table 1. Results for ComplexityNet (CNet). Model with clutter denoted with "-C", symmetry "-S" and clutter and symmetry with "-CS" suffix. UAE denotes the 'Unsupervised Activation Energy' [13]. Score performance measured with Pearsons correlation, map performance measured by Pearsons 2D correlation following prior work [14, 18].

3.3. Prediction Results

We show the results of our ComplexityNet architecture in Table 1. Results for ComplexityNet are cross-validated, taking the average performance over an 8-fold split of the training data. We compare our architecture against the previous state-of-the-art approach, the Unsupervised Activation Energy (UAE) [13], which is also the only prior approach that also generates "complexity maps" which we can compare against. Our results are shown in Table 1, and examples of images and ComplexityNet predictions are given in Figure 3. We find that our approach outperforms the UAE method (an increase of 80%) when considering predicting single-score ratings for the VISC-C dataset. We also find that we are able to generate complexity maps that better match human data than the comparative unsupervised method. While clutter alone offers the best single-score performance, only with clutter and symmetry combined does the network perform best at region prediction. Our approach allows us to predict not just the complex regions of scenes, but also the areas that humans find 'simple'; which may explain the improved performance over prior work that only examines the 'complex' regions. We show some examples of UAE maps against ground-truth human complexity maps in Figure 4. It can be seen that UAE maps highlight some regions of the image as complex that humans do not find complex.

Our original hypothesis, based on prior work in the cognitive science, was that clutter and symmetry may play a role in how humans perceive the complexity of natural scenes. Our results provide additional evidence that this is the case; by including auxilliary clutter and symmetry inputs into our architecture we boost score prediction performance by over 3%, a small but considerable increase considering the very small dataset size. Interestingly, clutter and symmetry provide no benefit to predicting where the *complex* regions of scenes lie; but appear to be more important for learning which regions of a scene are perceptually simple to humans.

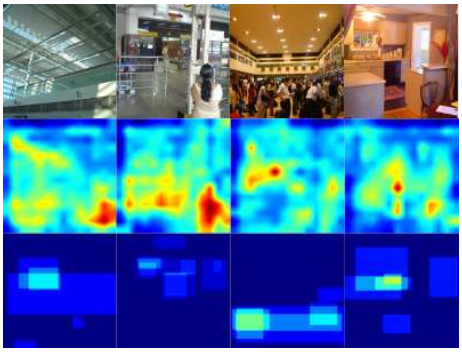

Fig. 4. Comparison between UAE maps [13] (middle row) against ground-truth human data (bottom row) that indicates 'complex regions'.

4. CONCLUSION

In this paper we propose a novel deep learning model for perceptual complexity estimation; predicting the regions of natural scene images that humans find complex or simple. The proposed model, ComplexityNet, makes use of clutter and symmetry metrics, evidenced to relate to human perception of scene complexity. We find that our metrics do indeed provide a performance boost, and our proposed approach exceeds the accuracy of the previous state of the art method over a dataset of 800 scene images, for both score and map prediction. We are capable of predicting the scores humans give scenes, when asked to rate complexity numerically, with a Spearmans correlation of 0.716 and can reproduce complexity maps with good accuracy for both simple and complex regions of images. There is much left to be explored with determining how humans process the complexity of images. Available datasets are limited, and could be expanded, and there is little data on how the complexity of scenes influences other perceptual characteristics of those scenes; such as how easily they are remembered. Future work may use models of complexity to examine these characteristics in other datasets without requiring difficult, and expensive, ground-truth data gathering.

5. REFERENCES

[1] George David Birkhoff, *Aesthetic measure*, Harvard University Press, 1933.

[2] Joan G Snodgrass and Mary Vanderwart, "A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity.," *Journal of experimental psychology: Human learning and memory*, vol. 6, no. 2, pp. 174, 1980.

[3] Christopher Heaps and Stephen Handel, "Similarity and features of natural textures.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 25, no. 2, pp. 299, 1999.

[4] Aude Olivia, Michael L Mack, Mochan Shrestha, and Angela Peeper, "Identifying the perceptual dimensions of visual complexity of scenes," in *Proceedings of the annual meeting of the cognitive science society*, 2004, vol. 26.

[5] Honghai Yu and Stefan Winkler, "Image complexity and spatial information," in *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, 2013, pp. 12–17.

[6] Maurizio Cardaci, Vito Di Gesù, Maria Petrou, and Marco Elio Tabacchi, "A fuzzy approach to the evaluation of image complexity," *Fuzzy Sets and Systems*, vol. 160, no. 10, pp. 1474–1484, 2009.

[7] Andrei N Kolmogorov, "Three approaches to the quantitative definition ofinformation'," *Problems of information transmission*, vol. 1, no. 1, pp. 1–7, 1965.

[8] Ruth Rosenholtz, Yuanzhen Li, and Lisa Nakano, "Measuring visual clutter," *Journal of vision*, vol. 7, no. 2, pp. 17–17, 2007.

[9] Jaume Rigau, Miquel Feixas, and Mateu Sbert, "Conceptualizing birkhoff's aesthetic measure using shannon entropy and kolmogorov complexity.," in *Computational Aesthetics*, 2007, pp. 105–112.

[10] Silvia Elena Corchs, Gianluigi Ciocca, Emanuela Bricolo, and Francesca Gasparini, "Predicting complexity perception of real world images," *PloS one*, vol. 11, no. 6, pp. e0157986, 2016.

[11] Silvia Corchs, Gianluigi Ciocca, and Francesca Gasparini, "Human perception of image complexity: real scenes versus texture patches," in *Journal of Alzheimer's Disease*, 2016, vol. 53, p. s51, Abstracts for the Second International Meeting of the Milan Center for Neuroscience (Neuromi): Prediction and Prevention of Dementia: New Hope (Milan, July 6–8, 2016).

[12] Fintan Nagle and Nilli Lavie, "Predicting human complexity perception of real-world scenes," *Royal Society Open Science*, vol. 7, no. 5, pp. 191487, 2020.

[13] Elham Saraee, Mona Jalal, and Margrit Betke, "Visual complexity analysis using deep intermediate-layer features," *Computer Vision and Image Understanding*, vol. 195, pp. 102949, 2020.

[14] Erdem Akagunduz, Adrian G Bors, and Karla K Evans, "Defining image memorability using the visual memory schema," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 9, pp. 2165–2178, 2019.

[15] Cameron Kyle-Davidson, Adrian G Bors, and Karla K Evans, "Modulating human memory for complex scenes with artificially generated images," *Scientific Reports*, vol. 12, no. 1, pp. 1–15, 2022.

[16] "VISC-C: https://ccpl.hosted.york.ac.uk/research," .

[17] Jianbo Shi and Jitendra Malik, "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[18] Cameron Kyle-Davidson, Adrian Bors, and Karla Evans, "Predicting visual memory schemas with variational autoencoders," *British Machine Vision Conference, BMVC*, 2019.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Identity mappings in deep residual networks," in *Proc. European conference on computer vision*. Springer vol. LNCS 9908, 2016, pp. 630–645.

[20] 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*, 2016, pp. 770–778.

[21] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., pp. 8024–8035. Curran Associates, Inc., 2019.