Visual Taxometric Approach to Image Segmentation Using Fuzzy-Spatial Taxon Cut Yields Contextually Relevant Regions

Lauren Barghout

Berkeley Institute for Soft Computing (BISC),
U.C. Berkeley, California, United States
lauren.barghout@gmail.com
http://www.laurenbarghout.org

Abstract. Images convey multiple meanings that depend on the context in which the viewer perceptually organizes the scene. By assuming a standardized natural-scene-perception-taxonomy comprised of a hierarchy of nested spatial-taxons [17] [6] [5], image segmentation is operationalized into a series of two-class inferences. Each inference determines the optimal spatial-taxon region, partitioning a scene into a foreground, subject and salient objects and/or sub-objects. I demonstrate the results of a fuzzy-logic-natural-vision-processing engine that implements this novel approach. The engine uses fuzzy-logic inference to simulate low-level visual processes and a few rules of figure-ground perceptual organization. Allowed spatial-taxons must conform to a set of "meaningfulness" cues, as specified by a generic scene-type. The engine was tested on 70 real images composed of three "generic scene-types", each of which required a different combination of the perceptual organization rules built into our model. Five human subjects rated image-segmentation quality on a scale from 1 to 5 (5 being the best). The majority of generic-scene-type image segmentations received a score of 4 or 5 (very good, perfect). ROC plots show that this engine performs better than normalized-cut [9] on generic-scene type images.

Keywords: visual taxometrics, natural vision processing, image segmentation, spatial taxon cut, fuzzy filter, spatial taxons, scene architecture, scene perception, fuzzy perceptual inference, fuzzy logic, image processing, graph partitioning.

1 Introduction

Segmenting images into meaningful regions is pre-requisite to solving most computer vision interpretation problems. Yet region relevancy depends less on the numeric information stored at each pixel, then on the computer vision task and corresponding scene architecture required to perceptually organize the constituent visual components necessary for the task. This presents a problem for automated image segmentation, because it adds uncertainty to the process of selecting which pixels to include or not include within a segment.

A. Laurent et al. (Eds.): IPMU 2014, Part II, CCIS 443, pp. 163–173, 2014.

An analogous problem exists for text document interpretation. Segmentation[1] of the document into its relevant components, such as characters, words, sentences or paragraphs, is pre-requisite to interpretation. However unlike images, text-documents have a standardized architecture with components designated by punctuation. Traditional punctuation and modern innovations such as hypertext[2] mark-up language (html), minimize uncertainty in the process of selecting which characters to include or not include within a segment.

Standardized architecture for written documents provide an example of a complex system that has proven to be stable across history, culture and technical innovation. As pointed out by Nobel Laureate Herbert Simon [11], "hierarchy is one of the central structural schemes that the architect of complexity uses". He further observes that "hierarchic systems have some common properties that are independent of their specific content" and he roughly defines a complex system as a system in which "the whole is more than the sum of the parts... in the pragmatic sense that given the properties of the parts and the laws of their interaction, it is not a trivial matter to infer the properties of the whole."

Text-document architecture succeeds because its structure is independent of content semantics. Letters, words, sentences and paragraphs follow the same structure regardless of whether they belong to a document discussing fashion, religion or nature.

The standardized natural-scene-perception-architecture described in this paper mimics text-document architecture in several ways: it's structured as a nested taxonomy, scene segment structure is independent of scene content semantics, standardized structure is used to minimize uncertainty as to which pixels belong within a segment; and architecture enables interpretation by delivering visually relevant components.

1.1 Visual Taxometrics and Spatial Taxons

Visual-taxometrics seeks to distinguish categorical visual percepts -such as figure/ground perception, from continuous visual percepts - such as distance or size. Spatial-taxons, categorical variables of 'whole things' such as foreground, object groups or objects (Barghout 2009), are 'building blocks' of scenes. In essence they serve as a proxy for the figural status of the region. When human subjects are asked to mark the center of the subject of the image, they tend choose the center of a spatial taxon with little variance and rarely choose locations defined solely by continuous visual percepts [17] [6]. Furthermore, evidence suggests that the frequency at which people choose spatial-taxons at a particular abstraction level, follow rank-frequency distribution similar to Zipf's law – independent of image content [6]. This is consistent with the law of least effort found in other cognitive systems and with Simon's observations of complex systems.

[1] Usually the literature uses the term 'parsing' instead of 'segmentation' to refer to breaking language into constituent parts. I chose this phrase to illustrate the information-normic (similarity) between image and language parsing.

[2] My collaboration with Roger Gregory, who pioneered hypertext with its inventor Ted Nelson, informed my understanding of this point.

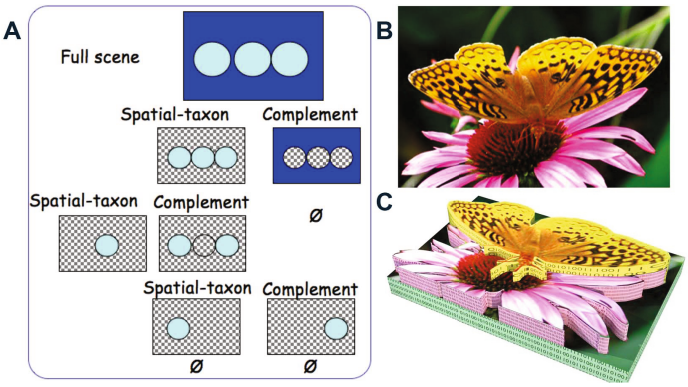

Fig. 1. (*A*) Natural-scene-perception-taxonomy comprised of a hierarchy of nested spatial-taxons. By assuming the taxonomy prior to segmentation, segmentation becomes a series of two-class fuzzy inferences. The full scene, top row, is at the highest level of abstraction. Each subsequent row is at lower level of abstraction within the taxonomy. (*B*) An image of a butterfly on a daisy. (*C*) A 3-dimensional version of image B where the third dimension (height) designates the abstraction level of the spatial taxon as shown in C.

The spatial-taxon view of scene perception assumes that humans parse scenes not between regions of similar features that vary continuously, but instead via discrete spatial 'jumps' biased toward taxometric scene configurations. Theories of visual attention make a similar distinction. The "spotlight theory" [12] assumes that attention regions vary continuously. Theories of "object based" attention assume that attended spatial regions vary in discrete location jumps as it accommodates attended objects.

If humans are parsing scenes by inferring categories, then quantifying pixel-region as to their aggregate "trueness" relative to the category prototype is prerequisite to human inspired computerized image segmentation. Humans assign meaning to visual percepts that they use to infer categories. Fuzzy-logic, which provides tools for handling partial or relative truth of meaning [15], enables inference based on visual percepts [4]. I've coined the phrase "natural-vision-processing" to refer to the parsing of images into psychological variables whose relative truth (fuzzy membership) corresponds to human phenomenological interpretation. Gestalt psychological variables such as similarity, good continuation, symmetry and proximity as first introduced by Wertheimer [13] provide the basis for fitting membership functions. A more detailed discussion on fitting Gestalt variables with fuzzy membership functions can be found in Barghout (2003) [4]. This paper focuses on fuzzy methods for optimizing spatial-taxon inference after a hypothetical set has been posited from Gestalt variables.

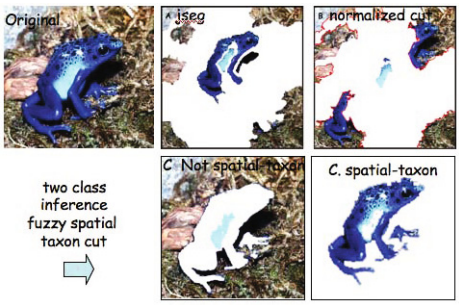

Fig. 2. Image segmentation algorithms, such as jseg and normalized-cut, produce what I call "jig-saw puzzle segments" of the original image (top left). In other words, though they do a good job of delineating regions of similar percepts, they are not meaningful to people. (A) Output of jseg algorithm [16]. From UCSB and downloaded 2010 ,version 6b. (B) Output of normalized-cut algorithm [18]. 2010 version. (C) The natural image processing engine output for not-spatial taxon inference and (D) spatial taxon inference.

1.2 Prior Work on Image segmentation

Most image segmentation algorithms stem from the school of thought that attention varies continuously over retinotopic location. Thus it makes sense to view images as a graph, image segmentation as a graph partitioning problem and precise high-dimension descriptive data at each graph node as pre-requisite to solving computer vision problems. For these approaches the criterion for graph partition is vital. They tend choose criterion of maximal contrast, where contrast is defined between summary statistics aggregated over candidate regions [9], [16]. Shi and Malik [9]) provide an excellent review of these methods. Though these methods succeed in parsing dissimilar regions, the regions in and of themselves are not meaningful. For example, figure 2 shows regions parsed to maximize differences between regions. The segments look like jigsaw puzzle pieces. Each jigsaw segment is not relevant to the visual understanding of the context and content or scene organization.

Fuzzy logic, however, provides an alternative school of thought where it makes sense to view images as spatially overlapping universe of discourses, image segmentation as a fuzzy set classification inference problem and the relative truth of the meaning of underlying a segmentation query [14] as pre-requisite to solving computer vision problems. In this way, image segmentation becomes a series of fuzzy two-class inference problems.

2 Fuzzy Natural Vision Processing and Spatial-Taxon Cut

By assuming the taxonomy prior to segmentation, parsing an image becomes a series of two-class fuzzy inferences. In this section, I will describe a system

that implements image segmentation as a nested two-class fuzzy inference system. Figure 3 provides an overview of the whole system. The sub-system (box A), shown on the left is similar to other fuzzy systems. It contains a fuzzification phase where the crisp values contained in the original image are re-parameterized into fuzzy cognitively relevant variables (CV). CVs are designed to fit human data or mimic human psychophysical and perceptual variables. A discussion along with detailed examples of calculations of fuzzy CVs can be found in Barghout (2003). Meaningfulness cues are composition styles with known CV spatial-taxon configurations. Its inference system, uses CV premises and meaningfulness rules to posit hypothetical spatial-taxons. Thus far, the fuzzy logic system is pretty standard in its design.

The next process (box C on the right), decides on the hypothetical spatial-taxon set and appropriate weighting. It is novel to this system. The system iterates through various combinations of hypothetical spatial-taxons, to infer the defuzzifed spatial-taxon that would result from each combination, and scoring the output for each combination. This enables posits to 'abstain'[3]. The score is a combination of spatial-taxon utility and the attentional resource requirement of the hypothetical spatial-taxon combination. The optimal set is chosen such that it maximizes utility and minimizes attentional resources.

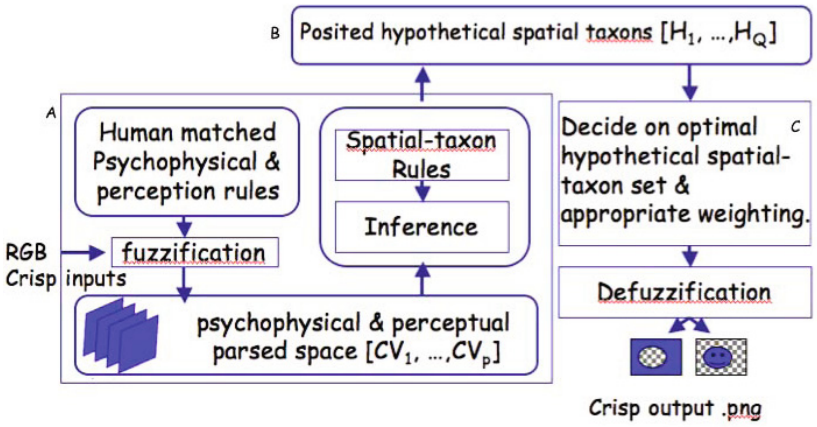

Fig. 3. Fuzzy natural-scene-perception system

The utility function I use to score the posited spatial-taxon was inspired by a seminal study of pictorial object naming [10] that found that objects were identified first at an "entry point" level of abstraction. Curious as to the whether the scene-architecture had an 'entry level' region, I undertook a multi-year study

[3] The idea to allow psychological detectors to abstain from contributing information to the system was suggested to me by Lotfi Zadeh in 2006, personal communication.

(2007-2011) surveying participants at the Burningman Arts Festival in NV, the Macworld conference in CA and the department of motor vehicles in Raleigh, N.C. The results suggest that images do indeed have an entry-level spatial taxon. Furthermore the spatial-taxon rank-frequency distribution measured in these studies suggest a law of least effort similar to that found in other cognitive processes [7]. Thus the utility function is inspired by the law of least effort. I define it operationally over an ordinal scale such that entry-level had the most utility, super-ordinate the next highest utility and all sub-ordinate decrease utility as a function of abstraction. This is a soft restriction, with granularity at abstraction levels. Use of attentional resources was also defined on an ordinal scale with granularity at the number of hypothetical spatial-taxons possible in the natural-vision processing engine. It's constrained to be inversely related to the number of significant spatial-taxon combination sets above threshold, where threshold was defined in terms of sub-population variance verses variance of the sub-population with the lowest within-group variance. This process is described in figure 4. Figure 5 provides a pictorial illustration using the image marked as "original" in figure 6.

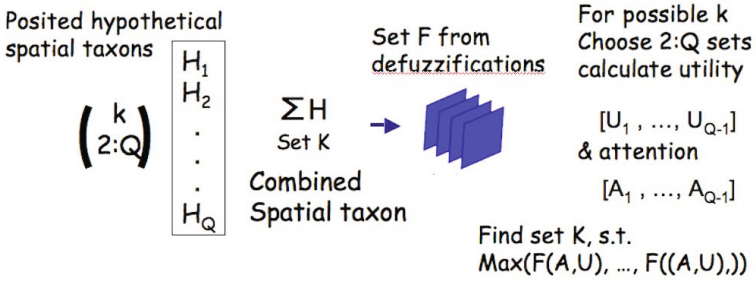

Fig. 4. Process description of box C in Figure 3

Partitioning by spatial-taxon cut has two phases. In the first phase we decide on the optimal hypothetical spatial-taxon set & appropriate rule weighting.

For $[^{k}_{2:Q}]$ defuzzifications calculate utility and attention-resources-requirement where

$$Utility(\Phi) = \int\int_{\Phi} hypothetical - spatial - taxon - utility(\Phi)d\Phi \qquad (1)$$

$$Attenional_{r}esources(\Phi) = \int\int_{\Phi} Attentional - inference - load(\Gamma)d\Phi \qquad (2)$$

Let A be a fuzzy set defined on a universe of Φ discrete meaningfulness cues $\Phi = [\Phi_1, \Phi_2, ..., \Phi_a]$ defined on the universe of discourse of two discrete scene

architecture states $S = [s_1, s_2]$ where s_1 is a spatial taxon and s_2 is the background[4] Set Φ represents the hypothetical spatial-taxons organizing constraints.

In the second phase, we "cut" the spatial taxon by defuzzifying the fuzzy conclusion. The crisp conclusion is normalized between zero and one. Spatial-taxon threshold is chosen according to use-case. In this system the threshold was set to 0.5.

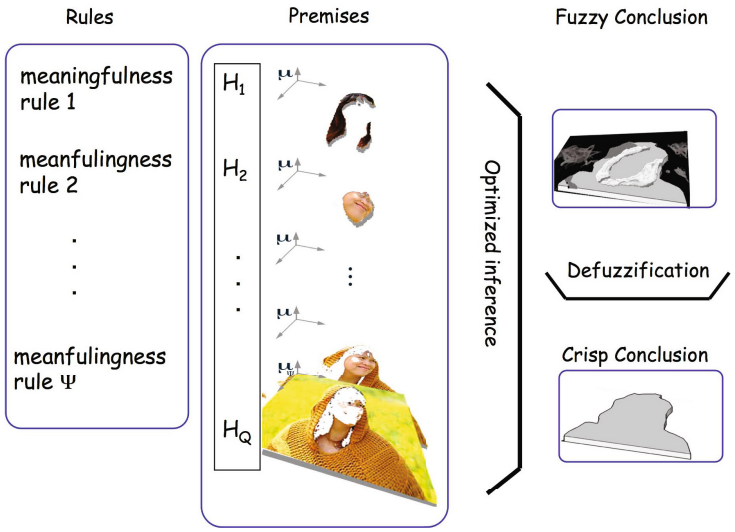

Fig. 5. Pictorial example of spatial-taxon inference as described in Figure 4. The mu axis designates the fuzzy firing power of each spatial taxon.

Figure 5, a woman wearing a hooded poncho in a field of yellow flowers, is used as an example. A series of meaningfulness cues are used to posit hypothetical spatial taxon. To make it easier to follow, I show cut-outs from the original image, next to the meaningfulness cues. An original is shown if the meaningfulness firing power of that pixel exceeds threshold. Note that because the poncho is orange, the intersection of yellow and red, it has membership in spatial-taxons and complement. These conflicting cues abstain because including them in the set drains attentional resources and provides little utility.

3 Performance Test Methods

70 real images composed of four "generic scene types", each of which required a different combination of the perceptual organization rules built into our model,

[4] Though the human perceptual state of "ground" extends beyond the subject and thus has fuzzy borders, it's digital image counterpart exists in a defined pixel set such that the "ground" is the complement of the spatial taxon.

Fig. 6. The Golden image was hand segmented by a human and is considered "ground truth". The Crisp Output is the spatial taxon inferred by the system. The difference between the Golden and system output is shown in Difference.

were collected. The natural-vision-processing system engine segmented them. Golden segmentations (ground truths) were manually segmented for each image using photoshop. A canny edge detector was used to produce the contours for both ground truths and system outputs. Fuzzy correspondence was calculated [3]. It was important to calculate fuzzy correspondence as opposed to crisp correspondence so as to not create error artifacts from slight offsets or registration errors. Hit-rate, false-alarm, correct-rejection and misses were determined and used to calculate ROC curves. The same procedure was used on a downloaded version of Normalized Cut [18].

Five human subjects rated image-segmentation quality on a scale from 1 to 5 (5 being the best). D-prime (detectability) was determined from the hit-rates and false alarm rates. Human subjects also rated the meaningfulness cues with results shown in Table 1.

Table 1. shows the four K spatial-taxon sets. An ANOVA was used to extract the relative proportion of meaningfulness cue for each corpus type - shown as linguistic hedges - as scored by 5 human subjects.

Cluster 1		**Cluster 2**	
Linguistic Hedge	Meaningfulness Cue	Linguistic Hedge	Meaningfulness Cue
abstain	Blurry	*some*	Blurry
some	Color Surround	*abstain*	Color Surround
very	Connected Taxon Color	*very*	Connected Taxon Color
abstain	Wall-like Background	*some*	Wall-like Background
Cluster 3		**Cluster 4**	
Linguistic Hedge	Meaningfulness Cue	Linguistic Hedge	Meaningfulness Cue
very	Blurry	*low*	Blurry
very	Color Surround	*some*	Color Surround
abstain	Connected Taxon Color	*some*	Connected Taxon Color
some	Wall-like Background	*not*	Wall-like Background

Fig. 7. Example engine outputs organized by meaningfulness cue combination cluster. In each example, the original image is on the left, the golden (hand segmented ground truth) in the middle and spatial taxon segmentation on the right.

4 Segmentation Results

ROC curves for all 70 images, figure 8a, show that the majority of images are well segmented. This is confirmed by humans scoring (5 subjects) that show that the majority of generic-scene-type images segmented via spatial taxon method received a score of 4 or 5 (very good, perfect). Figure 8b, ROC plots for 20 generic-scene-type images segmented using normalized cut and spatial taxon cut.

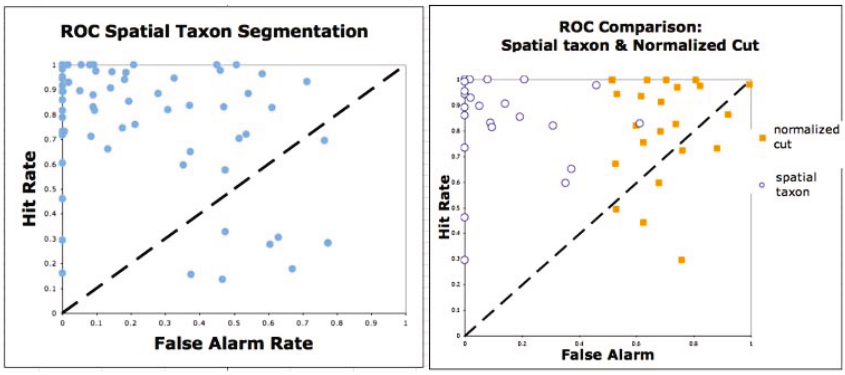

Fig. 8. (Left) ROC plot for spatial-taxon cut (70 images). (Right) Comparison between Spatial-taxon cut (circle) and normalized cut (square).

5 Conclusion

In conclusion, assuming taxonomy prior to segmentation enables quality parsing contextually relevant regions. A novel methodology that finds the optimal set and weight of premises performs well for optimizing spatial-taxon cut. Using fuzzy inference provides significant advantage for quantifying relative truth of a category, enabling cognitively relevant image segmentation. Both human grading and ROC plots show that this engine performs better than normalized-cut [9] on generic-scene type images.

Acknowledgments. Roger Gregory, Eyegorithm's co-founder and I co-wrote the code base on which the results were obtained. Special thanks to Dr. Christopher Tyler, Dr. Steve Palmer, Dr. Lotfi Zadeh and Dr. Lora Likova provide valuable feedback, suggestions and advice. Other collaborators include Haley Winter, Analucia DaSivla, Yurik Riegal, Colin Rhodes, Eric Rabinowitz, and Shawn Silverman. BurningEyeDeas LLC, an organization that does research at the Burningman art festival. Data posted at www.burningeyedeas.com.

References

1. Ruscio, J., Haslam, N., Ruscio, A.: Introduction To Taxometric Method. Lawrence Eelbaum Associates (2006)
2. Barghout, L.: Linguistic Image Label Incorporating Decision Relevant Perceptual, Semantic, and Relationships Data. USPTO. patent application 20080015843 (2007)
3. Barghout, L.: System and Method for Edge Detection in Image Processing and Recognition. WIPO Patent Application. WO/2007/044828 (2006)
4. Barghout, L., Lee, L.: Perceptual information processing system. USPTO patent application number: 20040059754 (2003)
5. Barghout, L., Sheynin, J.: Real-world scene perception and perceptual organization: Lessons from Computer Vision. Journal of Vision 13(9) (July 24, 2013)
6. Barghout, Winter, Riegal: Empirical Data on the Configural Architecture of Human Scene Perception and Linguistic Labels using Natural Images and Ambiguous figures. In: VSS 2011 (2011)
7. Cancho, Sole: Zipf's law and random texts. Advances in Complex Systems 5(1), 1–6 (2002)
8. James, W.: Principles of psychology, p. 403. Holt, New York (1890)
9. Shi, J., Malik, J.: Normalized Cuts and Image Segmentation. IEEE TPAMI 22(8) (2000)
10. Jolicoeur, Gluck, Kosslyn: Pictures and names: making the connection. Cognitive Psychology 16, 243–275 (1984)
11. Simon, H.: The Architecture of Complexity. Proceedings of the American Philosophical Society 106(6), 467–482 (1962)
12. Treisman, A.M.: Strategies and models of selective attention. Psychological Review 76(3), 282–299 (1969)
13. Wertheimer, M.: Laws of Organization in Perceptual Forms (partial translation). In: Ellis, W.B. (ed.) A Sourcebook of Gestalt Psychology, pp. 71–88. Harcourt Brace (1938)
14. Zadeh, L.: Outline of a new approach to the analysis of complex systems and decision processes. IEEE Trans. Syst. Man & Cybern. SMC-3 (1973)
15. Zadeh, L.: Toward a Restriction-centered Theory of Truth and Meaning (RCT). Information Sciences 248 (2013)
16. Deng, Y., Manjunath, B., Shin, H.: Color image segmentation. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2 (1999)
17. Barghout, L.: Empirical Data on the Configural Architecture of Human Scene Perception using Natural Image. J. Vis. 9(8), 964 (2009), doi:10.1167/9.8.964
18. Berkeley Segmentation Database, `http://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds`