Information Content Driven Unsupervised Top-Down Image Segmentation

Emanuel Diamant
VIDIA-mant, POB 933, Kiriat Ono 55100, Israel
emanl@012.net.il

Abstract

Image handling and processing is senseless if image information content is not taken into account. The essence of the latter is hard to define. Generally, we use it in the Shannon's sense, which means information content assessment averaged over the whole signal ensemble. However, humans rarely resort to such estimations. They are very effective in decomposing the viewed scene into its meaningful constituents and focusing the attention to the relevant scene parts. That means, performing reasonable image segmentation. We argue that information content definition in Kolmogorov's sense is more suitable for such cases. Following the concepts of the Kolmogorov's complexity theory, we propose to define image information content as a set of descriptions of the data structures discernible (segmentable) within an image. We propose a technique for creating such information content descriptions, which presumes a top-down unsupervised image segmentation process flow. We provide some illustrative examples, which demonstrate the effectiveness of our approach.

1. Introduction

Despite of the widespread use of digital imaging, the basic notion about what is visual information and what it essentially implies remain intuitive, uncertain and ambiguous. Most often, the expression "image information content" is used in the traditional Shannon's sense, which implies an average measure of uncertainty associated with the image generating process. Recently, it has become better appreciated that measuring the randomness of a picture does not capture its inherent structure, that is, the intricate correlations between its constituents. It became generally agreed that information content is more adequately represented by the measure of image complexity, which reflects the regularities present in an object.

At first, the idea to use complexity as a measure of information content was introduced (independently and approximately at the same time) by R. Solomonoff (1964) [1], A. Kolmogorov (1965) [2], and G. Chaitin (1966) [3]. But over the time, the name "Kolmogorov Complexity" has become far more widely mentioned. Following the theory of Kolmogorov's Complexity, image information content can be defined as a set of descriptions of discernable image data structures perceived at different visibility levels. As such, three perceptual (description) levels can be generally distinguished: 1) the global level, where the coarse structure of the entire scene is initially outlined; 2) the intermediate level, where structures of separate, non-overlapping image regions usually associated with individual scene objects are delineated; and 3) the low level descriptions, where local image structures observed in a limited and restricted field of view are resolved. Assuming that the descriptions are created with a syntactically defined and fixed language, the total length of the descriptors may be considered as a quantitative measure of the image contained information.

2. Creating information descriptors

Kolmogorov Complexity is a mathematical theory devised to explore the notion of randomness. Its basic concept is that information contained in a message (obviously, an image can be considered as a message) can be quantitatively expressed by the length of a program, that (when executed) faithfully reproduces the original message, [4]. Such a program is called the message description. Various description languages can be devised and put to use for the purpose of description creation, and it is only natural to anticipate that a specific language will influence the length of the description and its accuracy. One of the important findings provided by the Kolmogorov's complexity

theory is the notion of language invariance. That is, the chosen description language, of course, affects the length of object's description, but this influence can be taken into account by a language dependent constant added to the body of a language independent description, which actually is the Kolmogorov complexity of an object. The latter determines the absolute amount of information in an individual object, and thus can be called the absolute Kolmogorov complexity, [4]. The problem, however, is that this absolute Kolmogorov complexity is (theoretically) unconstrained and, thus, it is practically uncomputable.

A possible arrangement may be to give up in advance for an accurate information description, and to be satisfied with a less complete its version. Practically that means that some part of image information would remain undiscovered and undescribed. But essentially, we seldom use all the available information. Far more important for us is the insight of Kolmogorov complexity theory that in any case effective object description must commence with the simplest object structure delineation. An important equivalence between the shortest object description and the simplest object structure is established, [5]. The best way to achieve an object simplification is some sort of object compression, when the existing object regularities are simply squeezed out from it, [6]. The remaining part of the object data, the structure of which was not captured at the first stage, should be processed in the same manner – that is, the regularities discernable at some level must be squeezed out and the structures observable at this level must be described (encoded), [7]. A hierarchical and recursive strategy for a description creation is thus emerged: Beginning with the simplified and course object structure, the description is subsequently augmented with more and more fine details unveiled at different hierarchical levels of object analysis and description.

3. Relevant Background

Traditional approaches, which deal with information content descriptions (like the recently introduced MPEG-7 multimedia content description standard and its predecessor MPEG-4), begin information features gathering (for the purpose of information descriptors creation) in a quite different manner. They start with the low-level elementary information pieces gathering (the so-called bottom-up information processing approach) that are initially searched and retrieved over the entire image space. These pieces are then grouped and aggregated into larger and more complex agglomerations, which are fed to the higher system levels for farther generalization, classification, and other (higher-level) processing. To accommodate for external (user or system) requirements, that is, to incorporate the rules and principles by which disordered information pieces are combined and agglomerated, a supervised top-down control flow is generally assumed. Its aim is to mediate the bottom-up information gathering. It is generally believed that this supervised intervention of a top-down conscious control will lead to a more suitable and more task-fitting low-level information features acquisition, [8].

The roots of such preliminary bottom-up processing can be traced back to Treisman's Feature Integrating Theory [9] or Biederman's Recognition-by-components theory [10]. Relying on the evidence from human attentional vision studies, they were the first to propose the bottom-up manner of primary information gathering. Biological vision was always recognized as an unlimited source of inspiration for the computer vision designers. Endorsed by millions of years of natural evolution, this biological vision solution has become (since the early 80-s) an indispensable part of computer vision practice for many years to come.

However, the latest evidence from biological vision investigations, (especially from the field of selective visual attention that rules the eyes' saccadic movement), put the correctness of the traditional approach in doubts. To properly understand the point, some words must be spent to explain the matter: Human eye's retina has an odd and a bizarre structure – only a small fraction of its view field (approximately 2^0 out of the entire field of 140^0, [11]) is densely populated with photoreceptors. Just this small fragment of the retina (the so-called fovea) is responsible for our ability to see a sharp and clear picture of the surrounding world. The rest of the view field – is a fast descending (in spatial density) placement of photoreceptors, which provides the brain with a crude representation of the observed scene. To compensate for the lack of resolution over the entire visual field, continuous eye movements (also known as eye saccades) are performed, sequentially placing the high resolution fovea over various (information rich) scene locations.

According to attentional vision theories the decision to make a saccade and to fix the fovea over a new image location ***precedes*** high-resolution (low-level) image information gathering, and hence, it can be yielded only by the coarse information delivered by the peripheral vision. The flow of new evidence convincingly supports this suggestion: visual recognition/categorization tasks use "express", but comparatively imprecise and coarse-scale representations, before the fine-scale representations are acquired [12], the first signals

reaching the highest processing levels are from the eye's periphery, not from the fovea [13]. Not less surprising is the evidence that traditional assumptions about top-down intervention from the upper cognitive levels simply do not hold here. In most of the cases, saccadic movements are guided preattentively and unconsciously [14].

This flow of evidence from empirical studies of human attentional vision quite well support and come in agreement with the insights of Kolmogorov Complexity theory. Slightly twisted to fit the case of image information content exploration, the latter can be finally (and in brief) summarized as follows:

- Image information content is a set of descriptions of the observable image data structures.
- These descriptions are executable, that is, following them the meaningful part of image content can be faithfully reconstructed.
- These descriptions are hierarchical and recursive, that is, starting with a generalized and simplified description of image structure they proceed in a top-down fashion to more and more fine information details resolved at the lower description levels.
- Although the lower bound of description details is unattainable, that does not pose a problem because information content comprehension is generally fine details devoid.

4. Implementation Issues

Following the modern concepts of selective attention vision and the insights of Kolmogorov Complexity theory, we propose a new way for unsupervised top-down image segmentation facilitating meaningful information content revelation and gathering. Its architecture is shown in Figure 1, and it is comprised of three main processing paths: the bottom-up processing path, the top-down processing path and a stack where the discovered information content (the generated descriptions of it) are actually accumulated.

To facilitate the requirement for a top-down directed processing, we introduce a hierarchy of multi-level multi-resolution image representations called multi-stage image pyramid [15]. Such pyramid construction generates a set of compressed copies of the original input image. Each image in the sequence can be seen as an array that is half as large as its predecessor. The rules of this shrinking operation are very simple and fast: four non-overlapping neighbour pixels in an image at level L are averaged and the result is assigned to a pixel in a higher (L+1)-level image. This is known as "four children to one parent relationship".

At the top of the pyramid, the resulting coarse image undergoes a round of further simplification. Several image zones, representing perceptually discernible image fractions (visually dominated image parts, super-objects) are determined (segmented) and identified by assigning labels to each of the segmented pieces. Since the image size at the top is significantly reduced and since in the course of the bottom-up image squeezing a severe data averaging is attained, the image segmentation/classification procedure does not demand special computational resources. Thus, any well-known segmentation methodology will suffice. We use our own proprietary technique that is based on a low-level (local) information content evaluation [16].

The technique first outlines the borders of the principal image fragments. Then similarly appearing pixels within the borders are aggregated in compact spatially connected regional groups (clusters). Afterwards, every cluster is marked with a label. Thus, a map of labeled clusters, corresponding to perceptually discernible image regions, is produced. Finally, to accomplish top-level object identification, for each labeled region its characteristic intensity is computed as an average of labeled pixels. This way, a second (additional) segmentation map is produced, where regions are represented by their characteristic intensities.

From this point on, the top-down processing path is commenced. At each level, the two previously defined maps are expanded to the size of the image at the nearest lower level. The expansion rule is very simple: the value of each parent pixel is assigned to its four children in the corresponding lower-level map (a reversed shrinking operation). Since the regions at different hierarchical levels do not exhibit significant changes in their characteristic intensity, the majority of newly assigned pixels are determined in a sufficiently correct manner. Only pixels at region borders (and seeds of newly emerging regions) may significantly deviate from the assigned values. Taking the corresponding current-level image as a reference (the left side, bottom-up path belonging images), these pixels can be easily detected and subjected to a refinement cycle where they are allowed to adjust themselves to the "proper" nearest neighbors.

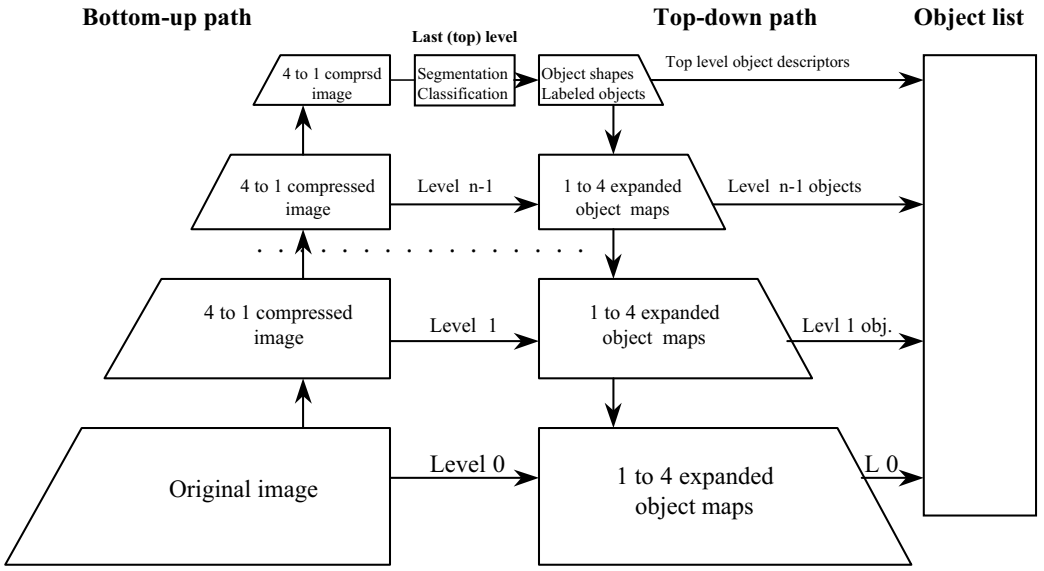

Fig. 1. The Block Diagram (Schema) of the suggested approach

In such a manner, the process is subsequently repeated at all descending levels until the segmentation/classification of the zero-level (original input image) is successfully accomplished. At every processing level, every image object/region (just recovered or an inherited one) is registered in the objects' appearance list, which is the third constituting part of the proposed scheme. The registered object parameters are the available simplified object's attributes, such as size, center of mass position (coordinates), average object intensity and hierarchical and topological relationship within and between the objects ("sub-part of…", "at the left of…", etc.). They are sparse, general, and yet specific enough to capture the object's characteristic features in a variety of descriptive forms.

This part of the processing scheme is (we suppose) the most suitable and natural place for external user interaction (a place for the "classical" top-down interference). User-defined task-dependent requirements can be easily formulated in human-friendly and human-accustomed forms, which are provided (supported) by the description implementations. The desired levels of description details are transparent (in the list) and are easily attended.

5. Experimental Results

To illustrate the qualities of the proposed approach we have chosen an image from the University of Washington Image Database [17], (Barcelona Image Collection).

Fig. 2 represents the original image, Figs. 3, 4, and 5 are examples of original image decomposition to regions of various scene complexity. For the given image size, the algorithm creates a six-level segmentation (pyramid) hierarchy. However, for the sake of space saving, we provide only several examples from this hierarchy (decomposition levels 5, 3, and 1), which for the reader's convenience are all expanded to the original image size. Extracted from the object list, the numbers of distinguished (segmented) at each level regions (objects) are given in each figure's capture.

Because in our approach the real objects are not known in advance, only intensity maps of growing complexity are presented here for the reason that they are perceptually close to the human's apprehension of image content.

Fig. 2. Original image, size 756x504 pixels.

Fig. 3. Level 5 segmentation, 24 objects (regions).

Fig. 4. Level 3 segmentation, 76 objects (regions)

Fig. 5. Level 1 segmentation, 250 objects (regions).

6. Conclusions

We presented a new technique for unsupervised image information content evaluation. Unlike traditional approaches, which adhere to bottom-up strategies, we propose a new scheme which produces the simplest (the shortest, in terms of Kolmogorov's Complexity) description of image information content. The level of unveiled description details is determined by the structures discernable in the image data and, thus, is independent from user intentions.

Despite a seeming similarity to the established multimedia content description standards, which (like MPEG-7 standard, e.g.) provide means and rules for image information content creation and Schemas for Object Description Design, our proposed approach is principally different: First, MPEG-7 description creation relies on a bottom-up process, [18]. This poses extreme difficulties for the initial object segmentation/identification. Therefore this task is left beyond the standard's scope. Secondary, MPEG-7 is not supposed to provide image reconstruction from the descriptions. Analogously designed descriptors can only be used for image comparison and similarity investigation purposes, (such as in Content Based Image Retrieval and other Web-related applications, [19]).

With respect to the standardized techniques, our approach has palpable advantages. We provide a technique that autonomously yields a reasonable image decomposition (to its constituent objects), accompanied by concise object descriptors that are sufficient for reverse object reconstruction with different levels of details.

References

[1] R. J. Solomonoff, "A Formal Theory of Inductive Inference", Part I, *Information and Control*, vol. 7, No. 1, pp. 1 – 22, March 1964.

[2] M. Li and P. Vitanyi, "An Introduction to Kolmogorov Complexity and Its Applications" (2nd ed), Springer-Verlag, New York, 1997.

[3] G. J. Chaitin, "Algorithmic Information Theory", *IBM Journal of Research and Development*, vol. 21, pp. 350-359, 1977.

[4] P. Grunwald and P. Vitanyi, "Kolmogorov Complexity and Information Theory", *Journal of Logic, Language and Information*, vol. 12, issue 4, pp. 497-529, 2003.

[5] N. Chater and P. Vitanyi, "Simplicity: A unifying principle in cognitive science?", *Trends in Cognitive Science*, vol. 7, issue 1, pp. 19-22, Jan. 2003.

[6] P. Vitanyi and Ming Li, "Minimum Description Length Induction, Bayesianism, and Kolmogorov Complexity", *IEEE Transactions on Information Theory*, vol. 46, No. 2, pp. 446 – 464, March 2000.

[7] N. Vereshchagin and P. Vitanyi, "Kolmogorov's Structure Functions with an Application to the Foundations of Model Selection", *Proceedings of The 43rd Annual IEEE Symposium on Foundations of Computer Science (FOCS'02),* pp. 751-760, Vancouver, Canada, Nov. 16-19, 2002.

[8] J. M. Wolfe, S. Butcher, C. Lee, and M. Hyle, "Changing your mind: On the contributions of top-down and bottom-up guidance in visual search for feature singletons", *Journal of Experimental Psychology: Human Perception and Performance*, vol. 29, pp. 483-502, 2003.

[9] A. Treisman and G. Gelade, "A feature-integration theory of attention", *Cognitive Psychology*, 12, pp. 97-136, Jan. 1980.

[10] I. Biederman, "Recognition-by-components: A theory of human image understanding", *Psychological Review*, vol. 94, No. 2, pp. 115-147, 1987.

[11] D. Van Essen, C. Anderson, D. Felleman, "Information Processing in the Primate Visual System: An Integrated Systems Perspective", *Science*, vol. 255, pp. 419-422, 24 Jan. 1992.

[12] A. Oliva and P. G. Schyns, "Coarse Blobs or Fine Edges? Evidence That Information Diagnosticity Changes the Perception of Complex Visual Stimuli ", *Cognitive Psychology*, v. 34, pp. 72 – 107, 1997.

[13] S. Thorpe, "Ultra-rapid scene categorization with a wave of spikes", *Biologically Motivated Computer Vision: Second International Workshop,* In: LNCS vol. 2525, pp. 1-15, Springer-Verlag, Berlin, 2002.

[14] R. Rao G. Zelinsky,, M. Hayhoe, D. Ballard, "Eye movements in iconic visual search", *Vision Research*, vol. 42, Issue 11, pp. 1447-1463, May 2002.

[15] S. Tanimoto and T. Pavlidis, "A hierarchical data structure for picture processing", *Computer Graphics and Image Processing*, Issue 4, pp. 104-119, 1975.

[16] E. Diamant, "Single Pixel Information Content", *Proceedings of SPIE-IS&T 15th Annual Symposium on Electronic Imaging*, SPIE vol. 5014, pp. 460-465, 2003.

[17] University of Washington Image Database http://www.cs.washington.edu/research/imagedatabase/...

[18] A. Jaimes and Shih-Fu Chang, "Automatic Selection of Visual Features and Classifiers", *IS&T/SPIE Conference on Storage and Retrieval for Images and Video Databases VIII,* San Jose, CA, Jan. 2000, SPIE vol. 3972.

[19] A. Jaimes and Shih-Fu Chang, "Model-Based Classification of Visual Information for Content-Based Retrieval", *IS&T/SPIE Conference on Storage and Retrieval for Images and Video Databases VII,* San Jose, CA, Jan. 1999.