Learning to Understand Image Content: Machine Learning Versus Machine Teaching Alternative

Emanuel Diamant

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

Abstract **Understanding image information content was always a critical issue in every image handling or processing task. Up to now, the need for it was met by human knowledge that a domain expert or a system supervisor have contributed to a given application task. The advent of the Internet has drastically changed this state of affairs. Internet sources of visual information are diffused and dispersed over the whole Web, so the duty of information content evaluation must be relegated now to an image content understanding machine or a computer-based program capable to perform image content evaluation at a distant image location. Development of Content Based Image Retrieval (CBIR) technologies is a natural move in the right direction. However... In this paper I will argue that the basic assumptions underpinning the majority of CBIR designs are wrong and inappropriate, (like many other basic conceptions that computer vision community proudly holds at this time).**

Index Terms— **Image understanding, image semantics recovery, knowledge acquisition and knowledge learning, image ontology creation and sharing.**

I. Introduction

The need for an explicit image information content comprehension was always recognized as a compulsory prerequisite for any image handling or processing task. Indeed, without clear understanding of what is contained in a given image, what kind of an anticipated image treatment can be seriously considered?

In the whole frame of "image information processing", the place of "image understanding" has been established more than 25 years ago by a famous bottom-up/top-down processing paradigm of Anne Treisman ([1]) and David Marr ([2]), which since then has not been subjected to any revision.

According to this paradigm, processing of image information content is understood as an interplay of two inversely directed processing streams. One is an unsupervised, bottom-up directed process of initial image information pieces discovery and localization (the so-called low-level processing stream). The other – is a supervised, top-down directed process, which conveys the rules and the high-level knowledge supposed to guide the linking and binding of the disjoint preliminary information pieces into perceptually meaningful image objects. (That is the high-level processing stream, associated with image understanding and cognitive image perception).

While the idea of low-level processing was always obvious and intuitively appealing (therefore, the mainstream of image processing even today is occupied mainly with low-level pixel-oriented computations), the issue of high-level processing from the very beginning was obscure, mysterious, and undefined. The paradigm says nothing about the roots of high-level knowledge origination or about the way it has to be incorporated into the introductory low-level processing. Until now, however, the problem was usually bypassed by capitalizing on the expert domain knowledge, adapted to each and every application case. It is not surprising, therefore, that the whole department of image processing has been (and continues to be) fragmented and segmented according to the high-level knowledge competence of domain experts. You can easily recall some obvious and well-known examples: medical imaging, aerospace imaging, infrared, biologic, underwater, geophysics, remote sensing, and so on "imagings".

The advent of the Internet, with huge volumes of information (including various forms of visual information) scattered over the web, raised an urgent demand for more general means of image semantics recovery, capable to handle visual information in a human-like intelligent manner and at a remote image location. However, deprived of any reasonable sources of a priori high-level image information, (and trapped by the tenets of bottom-up/top-down image processing), computer vision designers are forced to proceed only in one possible direction – they try to derive the needed high-level knowledge from the low-level information pieces.

Some theoretical work in biological and computer vision has been done to provide support and justification for such a development route. In this regard, two prevalent approaches are usually considered: chaotic attractor based approach [3], and saliency attention map based approach [4]. Both are computationally expensive. Both are presuming a Shannon-like sense of information, quite natural for a low-level bottom-up image processing arrangement. But all of them definitely violate the basic assumption about the supervisory role of high-level knowledge in the initial bottom-up processing. However, the pressure from Internet providers (and users) is extremely high. Internet-related interests undoubtedly dominate today the field of image semantics research and development. The marketing strategies are tough and impatient. Therefore, all contemporary Content-Based Image Retrieval techniques (currently in use or under development) unexceptionally attempt to derive high-level semantic information from the available low-level information details, (obviously, in a bottom-up proceeding manner), [5], [6].

0-7803-7724-9/03/$17.00 © 2003 IEEE

Another paradoxical example of high-level knowledge misapprehension and negligence can be drawn from the story of MPEG-4 (and higher number MPEG versions) standardization saga. Although the notion of image object (a high-level entity) has been indisputably declared as the fundamental item of image processing (mandatory for image segmentation and image coding processes), a clear definition of it (in the terms of high-level knowledge attributes) has never been suggested or introduced. Taking into account the entanglement of the bottom-up/top-down image-processing paradigm, it would be right to say that it will never be introduced in our time. As [7] has expressed this, it is "out of the scope of the standards and left to the content developers".

So, MPEG-4 standard, triumphantly announced as the first object-based encoding standard, has never been used for such critically important purposes, [8]. As well as the consequent MPEG-7 and MPEG-21 standards, which inherit their object rendering skills from the (declared, but not really existent) MPEG-4 properties.

So, millions of camera phones, produced by the industry in the last year, which were supposed to have MPEG-4 as their prime image-handling enabling technology, are forced to use its MPEG-2 rudimental option (which, obviously, does not have visual objects treatment facilities).

II. Challenging the State of the Art

Despite of many proud and loud declarations that we live today in an Information Age, any serious attempt to investigate traditional image processing from an information processing stand point had not happen yet (at least, to my humble knowledge). For that reason, several years ago, I have launched a preliminary exploration of this issue. The results were amazing and surprising [9]. It turns out that capitalizing upon the insights of Kolmogorov's complexity theory, one can prove that image physical information (information about visible image structures) can be derived (extracted) from an image without any high-level knowledge about it. That is, without any entanglement of high-level knowledge in the low-level image processing, without any interaction between high-level and low-level processing (streams). Obviously, that contradicts the traditional image-processing theories and the prevalent image-processing practice (all stemming from the same long-standing Treisman's paradigm). But it opens a unique opportunity to reconsider the traditional image processing approaches, to reject the obsolete ideas about low-level/high-level entanglement, and to start to investigate the two information processing modalities in an appropriate segregated manner.

Commencing such a move, it is quite natural to ask: if image physical information can be derived independently of the high-level knowledge about it, if high-level knowledge associated with an image is not a built-in property of an image, whose property is it, at last? The answer can be only one: this knowledge is a property of a human observer that is watching on an image! ("Beauty is in the beholder's eyes" – this truth was known to Baruch Spinoza more than 300 years ago). Today, we can only reinforce his insight – high-level image information (image semantics) is a product of human's brain, and all semantic judgements, all consequent image-handling decisions or image-inspired behavioral actions are an exclusive property and a privilege of a human observer.

What immediately follows from this is that the whole problem of high-level knowledge accommodation, representation, and management, which was traditionally associated with an image frame of references, must be now displaced and reallocated to the human frame of references. The latter is commonly known as the human's "ontology" (the human's understanding of the things in the surrounding world) – a huge and a complex knowledge base, which is gradually acquired and developed during the human life span. So, if we intend to facilitate a remote evaluation of image information content (by a computer-based program or a virtual machine, in short – a visual robot, as I would like to call it in the rest of the paper), we must provide it with something equal or equivalent to the human ontology.

What does it mean practically? While the right definition of ontology has not been established yet in the research community [10], – (the notion itself is very old, but its today's meaning is very different from that used by the ancient philosophers) – ontology is now an engineering issue, and a hot topic that is extensively studied and elaborated. From the vast spectrum of ontology definitions (and its implementation paradigms), we pick up the following one, provided by [12]:

"An ontology is a formal, explicit specification of a shared conceptualization.

Conceptualization refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon.

Explicit means that the type of concepts used and the constraints on their use are explicitly defined.

Formal refers to the fact that the ontology should be machine-readable.

Shared reflects the notion that an ontology captures consensual knowledge, that is, it is not private of some individual, but accepted by a group."

We hope that these concise definitions would help us to pave the way for a successful implementation of our plot.

III. Raising a Visual Robot Ontology

First of all, it must be stated unambiguously: the visual robot ontology, which we are intended to bring about, is not supposed to represent faithfully the entire world. It is supposed to work efficiently and serve the purposes for which it is deliberately appropriated. That allows to make it small enough, restricted, and concise. That does not invalidate the definitions just given above. On the contrary, that requires a more accurate and a more specific consideration of them.

Meeting the definition that concerns the shared nature of an ontology, we would have to remember that, unlike humans, the visual robot is not a self-governing system and is unable to develop its own ontology privately. So, aimed to create a

human-like visual robot that is able to interact and communicate with people, we must foresee a way to endow it with an ontology shared by such a bizarre community – a robot and its designer. Later this small community may be, of course, enlarged and extended, as additional members would be added and incorporated to it, humans and robots as well. In any case, for a successful teamwork, all participants would have to share a common ontology, which is lucid to them.

On the other hand, the options for a two-way communication exchange, so common in all other communities, in this case do not exist at all. We have here an unusual case, with a subordinate robot's position and a one-way form of robot/designer interrelationship. That creates a new and previously unknown paradigm of robot/designer collaboration in an ontology development, which will certainly influence the general system design philosophy. That means, the task of initial ontology creation, its relocation to robot's disposal, its unfolding and further development, all these sub-tasks become now a matter of special concerns of a human designer.

IV. Ontology Architecture

Thinking about the most appropriate ontology arrangement, it must be remembered that the gurus of the anticipated ontology-based designs have not presume that an application ontology must be (or will be) a functional equivalent of its human's counterpart. On the contrary, they assume that it will be small, modular, concise, and first of all, suitable and sufficient for the intended task accomplishment, [11], [12], [13]. In this regard, the expected ontology architecture seems to be a bundle of a limited (at least at the beginning) number of distinct ontology slots. Each single slot representing different aspects of a visual object description, different points of view on the object, different facets of object appearance or behavior. The slots can be created at different times, or at different development stages. Taking into account the gamut of features they are supposed to represent, the slots can be also called "partial" ontologies, [14].

To facilitate a concerted performance of partial ontologies bundle, a matching between conceptually similar parts of different ontology slots must be established. A vast literature on the subject does exist and could be found elsewhere, [15]. I will not stay on this too long. What I would like to emphasize, is that the prime duty of providing the appropriate matching rules must be committed to the system designer, (and not to be performed automatically, as it is usually attempted to be done in other design approaches). That is a very tricky and controversial question that desire a more careful contemplation, which is inevitably related to a more general question, the question of an appropriate mode of machine learning and its practical implementation.

V. The Learning Issues

The crucial issue of robot's learning and adaptation (in an unstable and changing surrounding) deserves a special consideration. The popular opinion that an intelligent system must be able to autonomously learn its changing and dynamic environment does not seem to me as an incontrovertible. Moreover, I think that the appropriateness of a machine learning approach is improperly exaggerated. After all, human ontology is an entangled mixture of different ontological forms that represent diversified levels of world understanding, that have been acquired at different stages of evolution development. Several categories of ontological descriptions could be distinguished in the human brain. There are ontological descriptions, evolutionary soldered into our DNA code. There are descriptions, which result from repetitive reinforcement learning, and descriptions acquired in an instructive and declarative manner.

Recent studies of human memory organization [16] and, correspondingly, memory learning "methodologies" reveal that several learning modalities coexist simultaneously in human brain. There is enough anatomical, physiological, and theoretical evidence to posit that the cerebellum is specialized for supervised learning, the basal ganglia are for reinforcement learning, and the cerebral cortex is for unsupervised learning [17]. It seems that humans in their everyday practice first of all resort to a supervised, declarative, and explanatory learning. New knowledge is acquired as explicit directives and illuminating statements that always come from the outside: from a parent to a child, from a tutor to a student, etc. (A first case of a teacher-pupil relationship in a non-human animal is recently reported in [18]). So, a declarative supervised mode of ontology creation and management, (in our case, a human-initiated, manually designed, by hand inserted and lined up specific-task-oriented ontology) is not an accidental oddity of this design but a carefully thought about and a deliberately chosen option.

VI. Matching Image and Its Ontology

The proposed structure of robot's ontology appears to be efficient also for solving the most crucial problem of contemporary computer vision – the problem of low-level/high-level image processing inconsistency. Indeed, the hierarchical structure of the image physical information description list, [9], closely resembles the hierarchical structure of a partial ontology. The latter is comprised of classes and sub-classes augmented by class attributes and rules, which determine the interrelations between them. In this regard, the description list can be seen as a partial ontology with undefined class labels but very well defined and abundant class and sub-class attributes. It is now the designer's duty to establish the appropriate mapping between these image attributes and the equivalent attributes of other partial ontologies.

By matching between attributes of a description list and equivalent attributes of other partial ontologies (in a bundle), class labels (concept names) could be reinstated in the physical image information ontology, thus fulfilling the desired image objects labeling. It must be especially emphasized: visual objects labeling with natural language names is the first, but the most crucial step of image semantics recovery. After all, semantics exists and can be comprehended only by words of a natural language. By establishing

equivalency between image objects and ontology classes (class labels or names), we enable farther semantic image analysis and exploration by climbing the ontology ladder in a usual text parsing manner.

VII. Some Concluding Remarks

The classical image-processing paradigm posits that image processing is an interplay of low-level and high-level information processing streams. The actual way of this interplay has been never defined well enough. Therefore, there is a widespread belief that the high-level semantics can be attained from the low-level processing stages.

In this paper, I am trying to reconsider this state of affairs. Capitalizing on the insights of Kolmogorov's complexity theory, I have shown that image physical information (information about visible image data structures) can be extracted from an image in an unsupervised top-down manner, totally independent from any high-level knowledge about image information content. What follows from this is: (1) High-level concepts are not involved in low-level image processing. (2) Semantics is not a property of an image, it is a property of a human observer watching on it. (3) The observer proceeds image semantics in accordance with his knowledge about the outer world, the so-called "world ontology". (4) A vision machine can be furnished with a replica of such an ontology, which does not need to be entirely full, but must be modular, concise, and specific enough about user's viewpoint on the task in hand. (5) Such a replica can not be get in a usual machine-learning fashion. It must be step-by-step established by a human supervisor – definitely a machine teaching endeavor.

I hope I was lucky to make my point clear enough.

References

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

[2] D. Marr, "Vision: A Computational Investigation into the Human Representation and Processing of Visual Information", Freeman, San Francisco, 1982.

[3] D.J. Amit, N. Brunel, "Learning internal representations in an attractor neural network with analogue neurons", Network, vol. 6, pp. 125-151, 1995.

[4] L. Itti, "Models of Bottom-Up Attention and Saliency", In: *Neurobiology of Attention*, (L. Itti, G. Rees, J. Tsotsos, Eds.), pp. 576-582, San Diego, CA: Elsevier, 2005.

[5] Xiang Sean Zhou, T.S. Huang, "CBIR: From Low-Level Features to High-Level Semantics", *Proceedings SPIE*, vol. 3974, pp. 426-431, San Jose, CA, January 24-28, 2000.

[6] C. Zhang and T. Chen, "From Low Level Features to High Level Semantics", In: *Handbook of Video Databases: Design and Applications*, by Furht, Borko/ Marques, Oge, Publisher: CRC Press, October 2003.

[7] S. Dasiopoulou, *et al*, "Knowledge-Assisted Semantic Video Object Detection", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, No. 10, pp. 1210-1223, October 2005.

[8] A. Puri, A. Eleftheriadis, "MPEG-4: An object-based multimedia coding standard", *Mobile Networks and Applications,* vol. 3, issue 1, pp. 5-32, 1998.

[9] E. Diamant, "Searching for image information content, its discovery, extraction, and representation", *Journal of Electronic Imaging*, vol. 14, issue 1, January-March 2005. Available: http://www.vidiamant.info.

[10] N. Guarino, "Understanding, Building, And Using Ontologies", *International Journal of Human and Computer Studies,* vol. 46, pp. 293-310, 1997.

[11] T.R. Gruber, "Toward Principles for the Design of Ontologies Used for Knowledge Sharing", In: *Formal Ontology in Conceptual Analysis and Knowledge Representation*, Kluwer Academic Publishers, 1993.

[12] A. Gomez-Perez, O. Corcho, M. Fernandez-Lopez, "Ontological Engineering (Advanced Information and Knowledge Processing)", Second Print, Springer Publisher, 2004.

[13] M. Uschold and M. Gruninger, "ONTOLOGIES: Principles, Methods and Applications", *Knowledge Engineering Review*, vol. 11, No. 2, pp. 93-155, 1996.

[14] P. Bouquet, F. Giunchiglia, F. van Harmelen, L. Serafini, and H. Stuckenschmidt, "C-OWL: Contextualizing Ontologies", *Second International Semantic Web Conference (ISWC-2003)*, LNCS vol. 2870, pp. 164-179, Springer Verlag, 2003.

[15] P. Shvaiko and J. Euzenat, "A Survey of Schema-based Matching Approaches", *Journal of Data Semantics*, 2005.

[16] L. Lin, R. Osan and J.Z. Tsien, "Organizing principles of real-time memory encoding: neural clique assemblies and universal neural codes", *Trends in Neuroscience*, vol. 29, No. 1, pp. 48-57, January 2006.

[17] K. Doya, "What are the computations of the cerebellum, the basal ganglia and the cerebral cortex?", Neural Networks, vol.12, issues 7-8, 11 October 1999, pp. 961-974.

[18] N. Franks, T. Richardson, "Teaching in tandem-running ants", *Nature*, 439, p. 153, 12 January 2006.