My Robot and I, We Both are Confident That We Have Already Seen This Picture

Emanuel Diamant
(emanl@012.net.il)

Extended Abstract: The remarkable success of the Internet and the rushing pace of WWW technologies have made enormous volumes of information suddenly accessible and available for all and everyone of us. Development of tools and techniques that would enable efficient search, browsing and management of this information became thus a task of paramount importance. However, while progress in processing textual information is certainly evident in the last years (recall, for example, the Google's success), processing of visual information remains a long-drawn gross confusion (a broh, in Yiddish). Aimed on creating smart machines that would possess human-like capability to grasp at a glance the image content, we still do not know how humans themselves perform their visual duties. Our understanding of the matter is framed by the famous low-level/high-level processing paradigm that Anne Treisman conceived more then twenty years ago. According to it, image information processing is an interaction of two inversely directed sub-processes. One is – an unsupervised, bottom-up evolving process of low-level elementary image information pieces discovery and localization. The other – is a supervised, top-down directed process, which conveys the rules and the knowledge that guide the assembling and linking of the elementary pieces just revealed into more large agglomerations and aggregations. It is generally believed that at some higher level of processing this interplay culminates with the required scene decomposition (segmentation) into its meaningful constituents (objects) that can be used for further image analysis and interpretation.

While the idea of low-level processing was always obvious and intuitively appealing (the mainstream of image processing for the most part even today is busy with low-level pixel-oriented computations), high-level processing from the very beginning was obscure, mysterious, and incomprehensible. To overcome the difficulties and to meet the growing need for high-level image clues, attempts to derive image semantics directly from the low-level image features are ubiquitously undertaken. It must be mentioned that all Content-Based Image Retrieval techniques (currently in use or under development) unexceptionally follow this principle.

In my talk, I will argue that this state of affairs is a delusion. In my recent research I have shown that image information content recovery is not an entangled interplay of two information-processing mechanisms, but a loose conjunction of them. In contrast to current views and practices, information about visible structures of image data (which I call "physical" image information) can be derived in a completely unsupervised manner, without any high-level knowledge intervention. Moreover, the top-down manner of such information recovery (the coarse-to-fine mode, as some people call it today) is the only right way to decompose an image into representative object sub-parts, which are suitable for further image content interpretation.

The dissociation between physical image information and its semantics opens a unique opportunity to investigate the nature of high-level image processing independently from the physical information and free from its interactive entanglements. In this regard, I will argue that semantics is not a property of an image, that images are endowed with semantics by a human observer, which is watching them. In humans, such an endowment is enabled by an extensive knowledge base that humans possess and develop during their life span. But in the case of an artificial vision system, which we are striving to create (a computer vision machine, a visual robot, etc.), we would have to provide it with something similar (to such a knowledge base). I will argue that furnishing my robot with such high-level knowledge (usually called "ontology") is feasible. It will force us to change our minds about some well-established dogmas related to the contemporary robot design approaches. Here is a list of critical points that must be immediately reconsidered:

- Ontology is a collective property, therefore means for sharing it between a robot and its designer (or between robot and other collaborating parties) must be foreseen and provided.
- Ontology is not an undividable whole, it can be seen as a set of particular modules, each one representing a specific problem aspect, and all together representing a partial knowledge about the world suitable for a task accomplishment.
- Ontology is an orchestrated modular structure. Adding, removing, changing or adjusting ontology modules, and maintaining their cooperative performance – all those things that are usually fused into one issue of ontology learning and adaptation – all them must be relocated to the designer's responsibilities. The common use of machine learning approaches is now improperly exaggerated. Supervised declarative learning is a natural and an inevitable solution.
- Naming image objects with corresponding ontology (class) labels is the right way for image semantics incorporation, because semantics is always carried out and is expressed only in linguistic forms, by words of a natural language.