ON THE POTENTIAL OF INCORPORATING KNOWLEDGE OF HUMAN VISUAL ATTENTION INTO CBIR SYSTEMS

Oge Marques and Liam M. Mayron

Department of Computer Science
and Engineering
Florida Atlantic University
Boca Raton, FL - 33431 - USA

Gustavo B. Borba and Humberto R. Gamba *

Programa de Pós-Graduação em Eng. Elétrica
e Informática Industrial
Universidade Tecnológica Federal do Paraná
Curitiba, PR - Brasil

ABSTRACT

Content-based image retrieval (CBIR) systems have been actively investigated over the past decade. Several existing CBIR prototypes claim to be designed based on perceptual characteristics of the human visual system, but even those who do are far from recognizing that they could benefit further by incorporating ongoing research in vision science. This paper explores the inclusion of human visual perception knowledge into the design and implementation of CBIR systems. Particularly, it addresses the latest developments in computational modeling of human visual attention. This fresh way of revisiting concepts in CBIR based on the latest findings and open questions in vision science research has the potential to overcome some of the challenges faced by CBIR systems.

1. INTRODUCTION

In this paper we consider recent developments in vision research – particularly visual attention – and how they apply to content-based image retrieval (CBIR). Our fundamental motivation is the realization that – in spite of more than 10 years of active research in this field – most CBIR research has primarily approached the problem from a preferred angle, particularly computer vision, using traditional techniques that have worked well in the past for problems in related domains. Since CBIR ultimately caters to the end user and the success of CBIR solutions hinges on capturing the essence of an image and how relevant it may be to a user's query, we postulate that a more detailed study of the latest vision research could lead to improved results.

2. BACKGROUND AND CONTEXT

2.1. Content-based image retrieval

CBIR is essentially different than the general image understanding problem. More specifically, it is usually sufficient that a CBIR system retrieves similar – in some user-defined sense – images, without fully interpreting its contents. CBIR provides a new framework and additional challenges for computer vision solutions, such as: the large data sets involved, the inadequacy of strong segmentation, the key role played by color, and the importance of extracting features and using similarity measures that strike a balance between invariance and discriminating power [1].

Ultimately, effective CBIR systems will have overcome two great challenges: the *sensory gap* and the *semantic gap*. The sensory gap is "the gap between the object in the world and the information in a (computational) description derived from a recording of that scene" [1]. The sensory gap is comparable to the general problem of vision: how one can make sense of a 3D scene (and its relevant objects) from (one of many) 2D projections of that scene. CBIR systems usually deal with this problem by eliminating unlikely hypotheses, much the same way as the human visual system (HVS) does, as suggested by Helmholz and its constructivist followers [2].

The semantic gap is "the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation" [1]. This problem has received an enormous amount of attention in the CBIR literature (see for example [3] and [4]) and is not the primary focus of the paper.

2.2. Vision Science

Vision science is an interdisciplinary field concerned with understanding how humans see, which considers phenomena of visual perception, the nature of optical information, and the physiology of the visual nervous system [2]. Vision science is a subset of cognitive sciences and shares many of the same concerns and interests. Of particular importance to this paper and the CBIR research community are the interdependent aspects of attention, perception, and memory [5].

Among the current research topics in vision science that may impact CBIR are:

- **Attention**: Is attention useful to CBIR as it has recently

*This research was partially sponsored by UOL (www.uol.com.br), through its *UOL Bolsa Pesquisa* program, process number 200503312101a.

been proven to be for object recognition? Can we factor the results of computational models of human visual attention into the design of CBIR systems? Which benefits can we expect from doing so? Which types of images or retrieval needs will be better addressed by this additional knowledge?

- **Perception**: CBIR systems must rely on each image's raw pixel values as if they contained the truth about how the image will be perceived by its viewers. Numerous experiments in vision science – ranging from optical illusions to inattentional blindness – show that this is hardly the case, and that the HVS, for the most part, cannot trust what it sees.

- **Memory**: Vision science research shows that the human brain builds different types of memory of visual events, depending on a number of factors, ranging from the duration of the stimuli to prior knowledge about the content of the scene, to the context in which it occurs. Moreover, the process of visual imagery – the building of mental images – maps closely to the way a CBIR user must behave in certain types of queries and therefore should be studied in more detail.

- **Contextual effects**: It has been proven that the perception of a scene or one of its components is strongly influenced by context information, ranging from recent stimuli to the expected position of an object within a scene. How much can CBIR systems learn about contextual influences and factor them into the system's design?

- **Function, category, language, and semantic meaning**: CBIR may benefit from looking at the roles played by perception of function and utility, perception of category, category prototypes, organization of categories into taxonomies or ontologies, and the role of language and cultural influences.

3. ATTENTION

There are many varieties of attention, but in this paper we are interested in what is usually known as *attention for perception*: the selection of a subset of information for further processing by another part of the information processing system. In the particular case of visual information, this can be translated as "looking at something to see what it is" [5].

It is not possible for the HVS to process an image entirely in parallel. Instead, our brain has the ability to prioritize the order the potentially most important points are attended to when presented with in a new scene. The result is that much of the visual information our eyes sense is discarded. Despite this we are able to quickly gain remarkable insight into a scene. The rapid series of movements the eyes make are known as *scanpaths* [6]. This ability to prioritize our attention is not only efficient, but critical to survival.

Broadbent states that cognitive processing, including attention, occurs sequentially [7]. Styles clarifies: when a new stimulus is presented attention is used to select the perceptual information that is stored in short-term memory. Later this may be moved to long-term memory [5].

There are two ways attention manifests itself. Bottom-up attention is rapid and involuntary. In general, bottom-up processing is motivated by the stimulus presented [5]. Our immediate reaction to a fast movement, bright color, or shiny surface is performed subconsciously. Features of a scene that influence where our bottom-up visual attention is directed are the first to be considered by the brain and include color, movement, and orientation, among others [8]. For example, we impulsively shift our attention to a flashing light. Complementing this is attention that occurs later, controlled by top-down knowledge – what we have learned and can recall. Top-down processing is initiated by memories and past experience [5]. Looking for a specific letter on a keyboard or the face of a friend in a crowd are tasks that rely on learned, top-down knowledge.

Both bottom-up and top-down factors contribute to how we choose to focus our attention. However, the extent of their interaction is still unclear. Unlike attention that is influenced by top-down knowledge, bottom-up attention is a consistent, nearly mechanical (but purely biological) process. In the absence of top-down knowledge, a bright red stop sign will instinctively appear to be more salient than a flat, gray road. Computational modeling of visual attention has made the most progress interpreting bottom-up factors that influence attention whereas the integration of top-down knowledge into these models remain an open problem.

3.1. Computational models of human visual attention

This section discusses several recently-proposed computational models of visual attention.

The Itti-Koch model of visual attention considers the task of attentional selection from a purely bottom-up perspective, although recent efforts have been made it incorporate top-down impulses [8]. The model generates a map of the most salient points in an image. Color, intensity, orientation, motion, and other features may be included as features. This map can be used in several ways. The most salient points can be extracted and individually inspected. Alternatively, the most salient regions can be segmented using region-growing techniques [9]. Another option is to use the most salient points as cues for identifying regions of interest [10]. Navalpakkam and Itti have begun to extend the Itti-Koch model to incorporate top-down knowledge by considering the features of a target object [11]. These features are used to bias the saliency map. In other words, if we want to find a red object in a scene the saliency map will be biased to consider red more than other features.

Draper et al. have shown that a simple implementation of visual attention (in their case, finding corners) can yield productive results [12] in the context of a CBIR system. They have modeled and implemented the *expert object recognition pathway*, the part of the brain that is responsible for recognizing specific objects.

Stentiford also uses a biologically-inspired model of visual attention for CBIR tasks [13]. It functions by suppressing areas of the image containing colors and shape that are repeated elsewhere. Flat surfaces and textures are suppressed while unique objects are given prominence. Regions are marked as high interest if they possess features not present elsewhere. The result is a visual attention map that is similar in function to the saliency map generated by Itti-Koch.

Machrouh and Tarroux have proposed using attention for interactive image exploration [14]. Their model uses past knowledge to modulate the saliency map to aid in object recognition.

Several other computational models of visual attention have been proposed and are described in [15].

3.2. Attention and similarity

Retrieval by similarity is a central concept in CBIR systems. Similarity is based on comparisons between several images. One of the biggest challenges in CBIR is that the user seeks semantic similarity but the CBIR system can only satisfy similarity based on physical features [1].

The notion of similarity varies depending on whether attentional resources have been allocated while looking at the image. Santini and Jain [16] distinguish *pre-attentive* similarity from *attentive* similarity: attentive similarity is determined after stimuli have been interpreted and classified, while pre-attentive similarity is determined without attempting to interpret the stimuli. They postulate that attentive similarity is limited to the recognition process while pre-attentive similarity is derived from image features [16].

Their work anticipated that pre-attentive (bottom-up) similarity would play an important role in general-purpose image databases before computational models of (bottom-up) visual attention such as the ones described in Section 3.1 were available. For specialized, restricted databases, on the other hand, the use of attentive similarity could still be considered adequate, because it would be equivalent to solving a more constrained recognition problem.

3.3. Attention, perception and context

Perception is sensory processing [5]. In terms of the visual system, perception occurs after the energy (light) that bombards the rods and cones in the eyes is encoded and sent to specialized areas of the brain. Perceptual information is used throughout to make important judgements about the safety of a scene, to identify an object, or to coordinate physical movements.

"Although the perceptual systems encode the environment around us, attention may be necessary for binding together the individual perceptual properties of an object such as its color, shape and location, and for selecting aspects of the environment for perceptual processes to act on" [5].

In a limited variety of tasks, such as determining the gist of a scene, perception can occur without attention [17]. However, for most other cases, attention is a critical first step in the process of perception.

Perception is not exclusively based on what we see. What we perceive is also a direct result of our knowledge and what we expect to see [18]. Many research studies have shown that the perception of a scene or the recognition of its components is strongly influenced by context information.

Two of the most notable ways by which the influence of context can be perceived are the influence of recent stimuli (*priming*) and the expected position of an object within a scene. In a classic experiment, Palmer [19] showed that participants were more successful in identifying objects when these were preceded by brief visual presentation of a context-appropriate scene, than if no scene was shown. For example, because of their shape similarities, when preceded by a picture of a kitchen table laid for breakfast, a US mailbox was identified as a loaf. In another classic experiment, Biederman [20] showed that subjects took longer to find a given target object in a randomly rearranged scene in which the object's positions did not match the users' expectations (e.g., a hydrant placed on top of a mailbox).

Even on a visual level (no semantics involved), the context in which an object of interest appears may influence our perception of it. An example is the simultaneous contrast effect – in which a gray object is perceived as darker or brighter depending on the surrounding background – and some of its variants (e.g., the Benary cross and White's illusion).

3.4. Global and local influences on attention

When we attend to a visual object we do so at different levels. Research by Navon [21] concluded that attention is directed to coarse-grained global properties of an object prior to analysis of fine-grained local details. A few years later, Stoffer [22] suggested that attention not only has to change spatial extent, but also has to change between representational levels. Stoffer suggests that the global level is usually attended to first but an additional step is required to reorient attention to the local level of representation.

4. CONCLUSION

The HVS is proof that in the majority of cases only a fraction of a scene must be attended to in order to make sense of the entire scene. CBIR systems, armed with the same facility, would be able to give more weight to salient parts of an image. To a human it would be both impractical and naive to study all parts of an image equally. Not only is the knowledge gained from studying non-salient features incremental, it distracts from the true meaning of the image.

CBIR systems will benefit by incorporating knowledge of human visual attention and perception. Attention can serve as a guide to portions of an image that require more specialized processing. Certain parts of an image contain more meaning than others. Using a model of perception would allow CBIR systems to analyze images in a fashion that is more intuitive to the user.

We expect that some of the ideas proposed in this paper will stimulate interest and awareness of the potential of incorporating knowledge of human visual perception research – which has been mostly overlooked and still has a number of open problems of its own – into the design of better CBIR systems.

In the future we expect that a further study of other components of vision science will yield similarly productive results. In particular, perception, memory, contextual effects, and semantic meaning hold promise for further study.

5. REFERENCES

[1] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. on PAMI*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000.

[2] S.E. Palmer, *Vision Science: Photons to Phenomenology*, MIT Press, Cambridge, MA, 1999.

[3] P. G. B. Enser and C. J. Sandom, "Towards a comprehensive survey of the semantic gap in visual image retrieval.," in *CIVR*, 2003, pp. 291–299.

[4] R. Zhao and W. I. Grosky, "Negotiating the semantic gap: from feature maps to semantic landscapes.," *Pattern Recognition*, vol. 35, no. 3, pp. 593–600, 2002.

[5] E. A. Styles, *Attention, Perception, and Memory: An Integrated Introduction*, Taylor & Francis Routledge, New York, NY, 2005.

[6] D. Noton and L. Stark, "Scanpaths in Eye Movements during Pattern Perception," *Science*, vol. 171, pp. 308–311, Jan. 1971.

[7] D. E. Broadbent, *Perception and communication*, Pergamon Press, New York, NY, 1958.

[8] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. on PAMI*, vol. 20, no. 11, pp. 1254–1259, Nov 1998.

[9] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004, pp. 11–37.

[10] O. Marques, L.M. Mayron, G.B. Borba, and H.R. Gamba, "Using visual attention to extract regions of interest in the context of image retrieval," in *Proceedings of the ACM SE'06*, Melbourne, FL, March 2006.

[11] V. Navalpakkam and L. Itti, "Modeling the influence of task on attention," *Vision Research*, vol. 45, no. 2, pp. 205–231, Jan 2005.

[12] B. Draper, K. Baek, and J. Boody, "Implementing the expert object recognition pathway," in *International Conference on Vision Systems, Graz, Austria*, 2003.

[13] F. Stentiford, "An attention based similarity measure with application to content based information retrieval," in *Proceedings of the Storage and Retrieval for Media Databases conference, SPIE Electronic Imaging*, Santa Clara, CA, 2003.

[14] J. Machrouh and P. Tarroux, "Attentional mechanisms for interactive image exploration," *EURASIP Journal of Applied Signal Processing*, vol. 14, pp. 2391–2396, 2005.

[15] L. Itti and C. Koch, "Computational modeling of visual attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194–203, Mar 2001.

[16] S. Santini and R. Jain, "The graphical specification of similarity queries," *Journal of Visual Languages and Computing*, vol. 7, no. 4, pp. 403–421, 1997.

[17] A. Oliva, "Gist of a scene," in *Neurobiology of Attention*, L. Itti, G. Rees, and J. Tsotsos, Eds., chapter 41. Academic Press, Elsevier, 2005.

[18] Z. W. Pylyshyn, *Seeing and Visualizing: It's Not What You Think*, MIT Press, Cambridge, MA, 2006.

[19] S.E. Palmer, "The effects of contextual scenes on the identification of objects," *Memory & Cognition*, vol. 3, no. 5, pp. 519–526, 1975.

[20] I. Biederman, "Perceiving real-world scenes," *Science*, vol. 177, no. 4043, pp. 77–80, 1972.

[21] D. Navon, "Forest before trees: The precedence of global features in visual perception," *Cognitive Psychology*, vol. 9, pp. 353–383, 1977.

[22] T.H. Stoffer, "The time course of attentional zooming: A comparison of voluntary and involuntary allocation of attention to the levels of compound stimuli," *Psychological Research*, vol. 56, pp. 14–25, 1983.