Image and Vision Computing 30 (2012) 472–473

Contents lists available at SciVerse ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Are we making real progress in computer vision today? ☆

Peter Meer

Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ 08854, USA

ARTICLE INFO

Article history:
Received 23 October 2011
Accepted 25 October 2011

Keywords:
Computer vision
Human vision

ABSTRACT

This paper presents an opinion on research progress in computer vision.

In short: probably not. To explore some of the challenges facing the field today, it helps to start with a thought experiment. Take two images, wherein the second image has many changes relative to the first; objects in the second image have altered illumination and exhibit completely different 3D poses. Concentrate on the correspondence between the points in these images: a point in the first image can be paired with a corresponding point in the second image, or simply has no matching pair due to their differences. Even when robust estimators assist in the recovery process, most of the correspondences amongst the points will not be valid. Though many of the points in the second image will be above a threshold, they will be in the wrong place. On the other hand, a human being is able to recognize this mapping immediately and with no trouble. Though the objects were rotated, the illumination was changed and many detected points in the second image were not in the first image, a human observer is able to interpolate the objects and compensate successfully for these differences.

So why does computer vision fail to execute this task effectively, while we as humans are able to excel at it with so little difficulty? Since birth, we have stored countless images in our brains, probably in a three-dimensional description. Creating an arsenal of three-dimensional images has helped us make sense of the world we inhabit. Once stored, these objects can easily be mentally rotated, assuming, of course, that the objects themselves are well-behaved; that is, they satisfy Gesalt-type laws.

It is then not a far leap to understand that we meld the bottom-up information we receive in a constant stream from our visual system with the top-down information stream we have stored over years of processing visual input. Our mental categories of objects encompass a variety of different forms, sizes or shapes. It is this accumulation of experiences that gives us the ability to comprehend and classify new images or observe altered ones. Unfortunately, scientists today have very little understanding of how images are represented in the brain.

Many researchers utilize today top-down information, but since the image classification is two-dimensional, top-down information is in a general sense just a projection of images onto a classification surface. Therefore, computers often cannot recognize large permutations of objects or that is the same object under different circumstances. Provide the computer with an untried database and utilize the same methodology: the result will generally be unsuccessful. It stands to reason that the classification executed by humans must operate differently.

Computer vision has more processing power and memory than in previous decades. Nevertheless, most of the successful algorithms (which can be applied broadly without failing), are at least fifteen years old, and Hartley and Zisserman's book on two and three-view geometry is over ten years old. Their text is probably the last that concentrates on some of the fundamentals. Newer books deal with more areas and applications, and often the applications are restricted to smaller domains. There is currently no push toward further developing the basic theories of computer vision.

Consider that most cameras nowadays can provide twelve-fourteen megapixels per image, which are at least 3500×3500 pixels. Relative to the retinal fovea, this is still only about one-fourth of the resolution. Moreover, in each moment the fovea focuses on a minuscule area. The scientists currently have no understanding of how a full image is synthesized from these small, focused areas, or how different resolutions in the periphery interact with fovea. But since the images have higher resolution, our low-level processing algorithms can yield gradual changes and can detect very small details when zoomed-in. Having only larger images and not newer, more reliable algorithms, is not going to solve the underlying problems.

Almost all constraints in computer vision are nonlinear and therefore should be processed with nonlinear algorithms, like that of Levenberg–Marquardt. These algorithms should be more widely taught and used. The top-down information processing of the humans should

☆ This paper has been recommended for acceptance by special issue Opinions Editor Sinisa Todorovic.

E-mail address: meer@cronos.rutgers.edu.

0262-8856
doi:10.1016/j.imavis.2011.10.004

appear also in computer vision and stored objects should be three-dimensional. Most orientations must come from these stored templates (and algorithms), not from the user. Our algorithms should be assessed more closely with visual phenomena which correlate with human perception. This will be a different kind of computer vision.

What we ultimately need is a paradigm shift. The term, coined in 1962 by philosopher Thomas Kuhn in "The Structure of Scientific Revolutions", is distinguished from normal science, in which scientists collect data in accordance with established theories without attempting to question the underlying assumptions. In computer vision today, researchers are solving small puzzles but moving away from truly challenging problems. The sooner we can recognize the need for changes also in the fundamentals, the sooner computer vision will begin moving in the direction of a paradigm shift.