Image Recognition Semantics: A Job for Smart Software or an Average Human

The marketing roar from vendors of semantic technology makes a complex indexing task easy and reliable.

Let's begin with a question: "Can an enterprise use software to figure out what a digital image or a video is "about"? ("About", as I am using the word, means looking at a snapshot of farm and recognizing the pigs, the cows, and chickens.)

Visualize your office building monitored by surveillance cameras. Instead of a human security guard watching for an intrusion, software "watches" the digital video and makes a decision about a specific individual attempting to enter the building.

The image recognition system plucks a person's face from the real-time video stream, matches it to a database, and determines whether he or she is a vice president or a stranger without access permission. The system "recognizes" the executive and unlocks the door.

For many years, security professionals have funded, tested, and tweaked commercial systems to make image recognition of faces a *reliable* reality, not a science fiction fantasy.

Alas, software sufficiently "smart" to figure out the identity of an individual or to determine the "aboutness" of a digital image is pot of gold long sought after but not yet found. One image recognition system I tested ingested a picture of a grade school class and after some digital cogitation displayed matching images of a Senate meeting. Basic "more like this" image processing remains a work in progress.

Google and Celebrity Facial Recognition

But advances are being made. In May 2011 Google's patent "Automatically Mining Person Models of Celebrities for Visual Search Applications" set off a flurry of commentary on blogs and mainstream publications like Forbes Magazine. US20110116690 was being downloaded when Google's chairman, Eric Schmidt, was explaining that image recognition was "too creepy". The story flashed from London to Mumbai. (See "Facial Recognition: Google Chairman Warns US Govt", May 20, 2011, at http://goo.gl/DPOuj. Google's back pedaling underscores the uncertainty about what a company should do with its image recognition technology, regardless of its capabilities. Was Google restricting technology because it came to close to what Eric Schmidt, Google's chairperson called, the "creepy line?"

Adding to the interesting synchronicity, Google, only days earlier, had updated its public Web Google Images' search function. A mouse click allowed me to sort results by subject or activate a color match. Google also displayed suggested image searches for "cage the elephant," "African elephant," "baby elephant", and so on. Not bad but not exactly a solution to the classroom-Senate committee match.

In my experience, computers struggle with snapshots. The ability of a computer to match the skills of a three year old pointing to an image in a scrapbook and saying, "That's my grandma" remains out of reach. In some of the organizations with which I am familiar, the mundane job of finding a specific image is still a manual process. And video? Even more difficult and made still more challenging even though the volume of rich media is exploding.

I will let you in on a secret. When I want some insight into next-generation search technology, I navigate to Google Research's "Publications by Googlers at http://research.google.com/pubs/papers.html." Although not a comprehensive archive, the technical papers provide a useful glimpse into search technologies from some of the world's most sophisticated engineers and scientists. In the category "Audio, Video, and Image Processing", there were more than 100 technical papers. last I checked.

I noted a number of suggestive research reports. These included "A Large-Scale Taxonomic Classification System for Web Based Videos" Yag Song, Ming Zhao, et al. Google's experts were testing a taxonomy with more than 1,000 categories. The idea was to use "smart software" to figure out what a video was "about". The Google method echoes Autonomy's approach, in my opinion, and demonstrated that Google algorithms can categorize video without metadata at an acceptable level of accuracy.

The 2009 article "Tree Detection from Aerial Imagery" by Lin Yang, Xiaqqing Wu, et al reveals that Google is working on figuring out the "what" in imagery. But the focus is on a quite narrow type of detection. I read this paper as an indication of a "building block" in a larger image processing capability. I took the same finding from "Face Tracking and Recognition with Visual Constraints in Real-World Videos", published by the IEEE in 2008. This paper complements "Large Scale Manifold Learning" (2008) by Sanjiv Kumar and Henry A. Rowley. The Google research explores performance, mathematical methods such as Markov algorithms, and variants of the Google PageRank algorithm. What is interesting is that Google, according to "Google Won't Release Awesome Facial Recognition App" http://www.pcworld.com/article/224007/google_wont_release_awesome_facial_recognition_app.html has potent image functionality that remains, for now, on the sidelines. Is this a due to a decision dictated by financial, legal, or technical factors? There is scant information about Google's plans for its image recognition technology. What is clear is that Google has invested time and effort in figuring out the content of static images and digital video. When Google does move, the impact on the market could be significant due to Google's near monopolistic control of search and retrieval.

Current Examples of What's Available

My view is that Google's consumer image search is useful, probably as good as, if not better than comparable systems from Bing (www.bing.com), TinEye (www.tineye.com), and Flickr (www.flickr.com). Personally I prefer the image search function of Exalead (www.exalead.com/image) which returns relevant images without the malware attracting iFrames used by Google. What few of my colleagues in the field of enterprise search know is that Dassault Exalead's system has for several years offered image search features only now becoming available on Google. For example, Exalead's system automatically recognizes an image suitable for desktop wall paper and displays a hot link to it. Exalead's portrait or

landscape option has been available for a long time. Exalead has also pushed ahead in video search. That system can be test driven at http://labs.exalead.com/experiments/voxalead.html and at http://www.exalead.com/search.

Autonomy also offers image and video search systems. Other vendors include such companies as OpenText's Nstein unit. Nstein uses technology from Imprezzeo (www.imprezzeo.com). The company employs content based image retrieval and facial recognition. Nstein's system has been tailored to the needs of those engaged in publishing. The user inputs or identifies a sample image. The system then displays matches. With some clicking, the result set can be narrowed to the image the user requires. Nstein provides a software development kit for the system.

A firm called IQ Engines offers an image recognition system which performs "computer vision search". You can try the technology at http://www.iqengines.com/. Click on Vision as a Service. A new window opens. You upload an image to the system. After of minute of processing, the system either displays matches or reports that the image was not in the database. My efforts to get the system to recognize an image of a shotgun wedding and runners passing a baton return null sets.

Kooaba is a visual recognition start up. The company offers a photo management system for licensees and an iPhone application. The user takes a picture of an object and uploads it to Kooaba. The system then "finds" similar images.

A key point is that these are systems are using metadata like the date, time, file type, and user generated description of an image. Algorithms create a "fingerprint" for color, shapes, and other discernable characteristics. If an image appears in a PowerPoint, the name of the PowerPoint "author" may be attached to the digital object. These systems are not figuring out whether the image is a prize winning heifer or a Volkswagen Jetta.

Image Recognition Applications

Confusion about image search, image recognition, and image systems is flourishing. One reason in my opinion is the failure to distinguish between the different applications to which image recognition can be applied.

Certain types of image processing work well and are well understood, and have a measureable impact. A good example is the machine vision sector of image recognition. Cognex (www.cognex.com) is one of the leaders in machine vision. The company's products make it possible to process barcodes for inventory control. Cognex's technology can "look at" a stream of manufactured components and "see" those with defects. (You may want to check out Orpix Computer Vision (http://www.orpix-inc.com), Pattern Recognition Company (http://www.pattern-recognition-company.com/machine-vision/optical-quality-control.html), and Microscan (http://www.microscan.com/en-us/technology/machinevisionsystems), among others.)

Cognex, despite the soft economy, has reported record revenue in its first quarter of 2011. The firm seems likely to push beyond $300 million in revenues. One indication of the

strength of this company is its cash position. The firm had a war chest in May 2011 or more than $300 million in cash and investment. At a time when traditional enterprise search vendors are struggling to stay afloat or tap investors for additional cash, Cognex is flying high.

There are some important differences between the image recognition needs in markets served by Cognex and the needs for image recognition in the part of marketing, sales, and business development people. A Cognex machine vision solution can be focused on a well defined domain, often with specific attributes or "tells". A defective chip, for example, may emit a different refractive index or have a discernible color variation. The technology to recognize a defect in a production line setting is extremely sophisticated. The return on investment can be calculated. Even at competitive labor rates, machine vision can pay for itself with speed, accuracy and at a lower cost than manual methods.

In marketing and sales, however, the person putting together a slide presentation needs an image of a product (relatively easy to find if there is metadata attached to the available pictures) or an image to show an intangible quality such as vigor (relatively hard even if someone has indexed an in-house image collection). Vendors offering image management systems based on metadata provided by the camera or by a human indexer are available. One can use the InMagic system as an image retrieval system. (http://www.inmagic.com/) Clever system administrators can make a traditional database like Oracle or SQL Server provide access to images.

But for larger collections of digital images—what used to be called 35 mm slide collections--one needs specialized DAM (digital asset management) systems from such vendors as Adobe, Canto, or Microsoft iView, among others. These systems offer version management, support for different image types such as Photoshop and PDF (portable document format), TIFF (tagged image file format), and vector drawing files. The systems include access controls, essential if an organization is doing work for certain government agencies. These systems focus on reducing bottlenecks in work flows. If a person needing an image cannot locate it, an Easter Egg hunt is required. Even with fancy systems, the amount of time required to find a specific image or a specific segment of digital video is indeterminate. Exalead's video search system does allow the user to view a video at the point at which the query matches the content of a digital video.

And What about Video?

Video can pose some additional challenges. Digital video is an unwieldy beast with an appetite for storage and a generous side dish of bandwidth. One company which has received accolades from industry groups and analysts is Altus Corp http://www.altuscorp.com/. The company offers on-demand rich media solutions for a range of enterprise applications. The Altus system can be used for knowledge sharing within an organization, a sales enabling service, an educational service, or a system to deliver video from a conference with multiple, simultaneous presentations.

Its flagship product is Vsearch. You enter key words into a search box. The system displays videos that match the query. and permits a user to assign ratings to them. A click in the results

list plays the presentation, audio file, or video. Like the Exalead system, Vsearch starts the video where the "hit" for the user's query appears in the file. The user can watch the complete video or just the section required. One useful feature is that from a results list, the user can download a PowerPoint, MP3 audio file, or an mp4 video to an iPhone, iPod, or other devices such as an Android mobile phone.

Altus has positioned itself as providing a service that "transforms enterprise video into a valuable asset for any organization. vSearch creates a cloud-based learning environment that combines enterprise video with PowerPoint slide synchronization and scrolling transcripts into an accessible video content archive that is searchable down to the spoken word or specific point of interest. Content can be viewed as streaming media or on-demand presentations from any computer, tablet or smart phone -- allowing instant access to knowledge anytime or anywhere." The Altus approach is to deliver video search as software as a service (SaaS). The firm's clients include Cisco Systems, General Motors, IBM, and Oracle, among others.

Net Net

Still, the question that interests me is, "Are these systems from sophisticated technology companies able to look at an image or a frame in the video and "figure out" what the picture represents?" The sci-fi version of image recognition is out of reach. Pictures can be about anything. The meaning of a picture depends on a context that, at this time, requires a human to discern. For now, humans still have a role to play in finding just the right image for any given situation. We are not about to see the end of the good old fashioned function called indexing for rich media for a few years.

What about semantics in enterprise image recognition technology? I think we should accept the American comedian Steve Martin's advice: "You know what your problem is, it's that you haven't seen enough movies. All of life's riddles are answered in the movies."

Stephen E Arnold, May 24, 2011

Mr. Arnold is a consultant. More information about his practice is available at www.arnoldit.com and in his Web log at www.arnoldit.com/wordpress.