Visual-Concept Search Solved?

Cees G.M. Snoek and Arnold W.M. Smeulders,
University of Amsterdam

Progress in visual-concept search suggests that machine understanding of images is within reach.

Interpreting the visual signal that enters the brain is an amazingly complex task, deeply rooted in life experience. Approximately half the brain is engaged in assigning a meaning to the incoming image, starting with the categorization of all visual concepts in the scene (S.E. Palmer, *Vision Science: Photons to Phenomenology*, MIT Press, 1999).

Nevertheless, during the past five years, the field of computer vision has made considerable progress. It has done so not on the basis of precise modeling of all encountered objects and scenes—that task would be too complex and exhaustive to execute—but on the basis of combining rich, sensory-invariant descriptions of all patches in the scene into semantic classes learned from a limited number of examples.

Research has reached the point where one part of the community suggests visual search is practically solved and progress has only been incremental (T.-S. Chua, "Towards the Next Plateau: Innovative Multimedia Research beyond TRECVID," *Proc. 15th Int'l Conf. Multimedia*, ACM Press, 2007, p. 1054), while another part argues that current solutions are weak and generalize poorly (J. Yang and A.G. Hauptmann, "(Un)Reliability of Video Concept Detection," *Proc. Int'l Conf. Image and Video Retrieval*, ACM Press, 2008, pp. 85-94). We've done an experiment to shed light on the issue.

BRIDGING THE SEMANTIC GAP

The general visual-retrieval problem is rooted in the *semantic gap*: the lack of correspondence between the low-level features that machines extract from a visual signal and a human's high-level conceptual interpretations. Researchers have proposed many solutions to bridge the gap—for example, by using text, speech, tags, or example images. However, the most cognitive approach is to type a concept from visual information and retrieve the images carrying that (automatically detected) concept, as Figure 1 shows.

The first step is to extract from an image locally measured features—lots of them, ranging from 40 to 100,000. The features are invariant descriptors that cancel out accidental circumstances of the recording caused by differences in lighting, viewpoint, or scale. To capture the world's complexity, many texture, shape, and color descriptors must be extracted.

The second step is to project the descriptors per pixel onto one of 4,000 words. These aren't real words, but rather summarizations of one local patch of the image describing a single detail: a corner, texture, or point. Researchers initially only summarized the image at the most salient points, but it now appears that full-density descriptions are superior.

In the third step, a machine-learning algorithm converts the visual words into one of the semantic concepts. In fact, it assigns a probability to all of the concepts simultaneously. These probabilities are used to rank images in terms of concept presence.

Researchers train the algorithm with the help of manually labeled examples. Because there are far more negative examples than positive ones, they intensively compute the optimal machine-learning parameters using grids and GPUs.

Detecting an object such as the American flag is relatively simple if the variance in sensory conditions such as illumination and shading can be accounted for—the flag always shows the same colors and color transitions. Note that a geometrical model of a flag would almost always fail, as flags rarely appear like straight squares. Detecting a walking person from one image requires a richness of poses learned from a labeled set. Snow is even harder to detect because it's white and texture-free and may assume all sorts of shapes.

 Published by the IEEE Computer Society

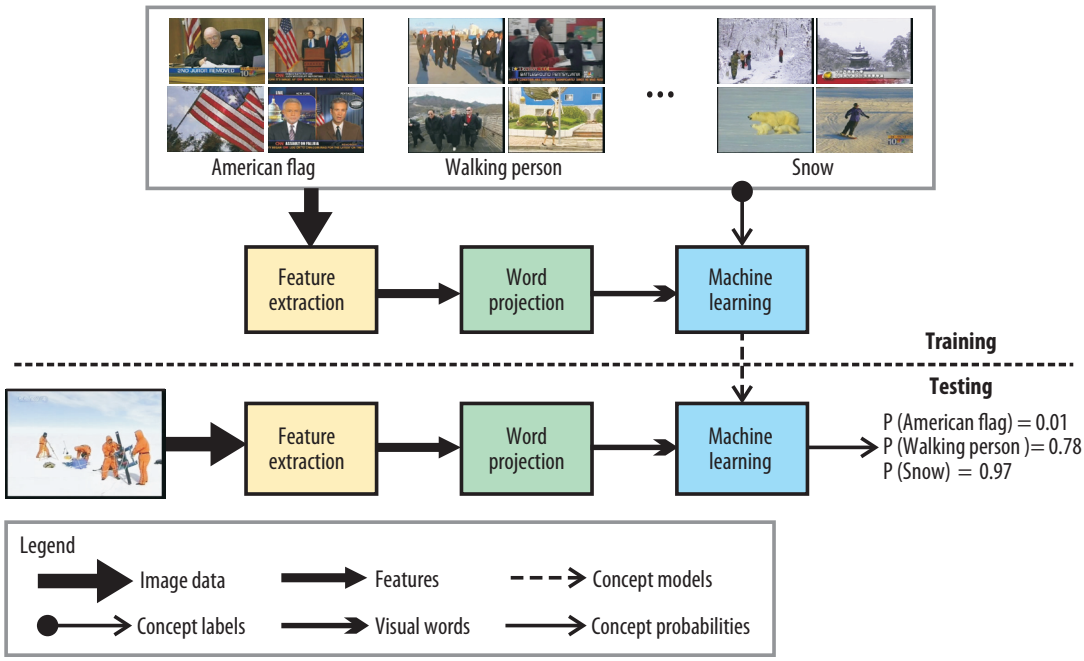

Figure 1. General scheme for detecting visual concepts in images, with three typical concepts highlighted. First, researchers project extracted image features into visual words. Then they train concept models from both the visual words and the concept labels using machine learning. Finally, during testing, researchers assign concept probabilities to previously unlabeled images.

Remarkably, although none of the features in current detection methods is specific to any of the concepts, the technique can still detect any of them with sufficient success.

SEARCH ENGINE BENCHMARKS

Crucial drivers for progress in visual-concept search are international search engine benchmarks such as ImageCLEF (Cross-Language Evaluation Forum), Pascal VOC (Visual Object Classes), PETS (Performance Evaluation of Tracking and Surveillance), and VACE (Video Analysis and Content Extraction). However, thus far the National Institute of Standards and Technology TRECVID (TREC Video Retrieval) benchmark has played the most significant role.

TRECVID aims to promote progress in content-based retrieval from digital video via open, metrics-based evaluation. With the support of 70 teams from academia and industry, including the University of Oxford, Tsinghua University, and IBM Research, it has become the de facto standard for evaluating visual-retrieval research. TRECVID has been an important driver for the community in sharing resources to validate visual-search experiments, most notably the manual annotations provided by the Large-Scale Concept Ontology for Multimedia (LSCOM).

Benchmarks' open character ensures the rapid convergence of effective concept-detection approaches. However, their reliance on relatively homogeneous training and test data has also prompted critics to suggest that overfitting has contributed to recent positive results. Given the community effort to push the envelope, it's fair to question the progress of visual-concept search.

MEASURING PROGRESS

We assessed image-categorization progress by comparing a state-of-the-art search engine from 2006 (C.G.M. Snoek et al., "The Challenge Problem for Automated Detection of 101 Semantic Concepts in Multimedia," *Proc. 14th Ann. ACM Int'l Conf. Multimedia*, ACM Press, 2006, pp. 421-430) with one from 2009 (C.G.M. Snoek et al., "The MediaMill TRECVID 2009 Semantic Video Search Engine," *Proc. 7th TRECVID Workshop*, NIST, 2009, www-nlpir.nist.gov/projects/tvpubs/tv9.papers/mediamill.pdf).

We used four mixtures of two broadcast video data sets obtained from the 2005 and 2007 TRECVID video-retrieval benchmarks. The first data set was from the MediaMill Challenge and included 85 hours of shot-segmented news video from China, Lebanon, and the US; the second was the training set of the TRECVID 2007 benchmark and contained 56 hours of shot-segmented Dutch documentary video. We separated both video data sets into an independent training (70 percent) and test (30 percent) set.

In our experiment, we used the two search engines to detect the most common visual concepts in the literature—namely, the 36 defined in LSCOM, as labeled manually for both data sets. We took into account both

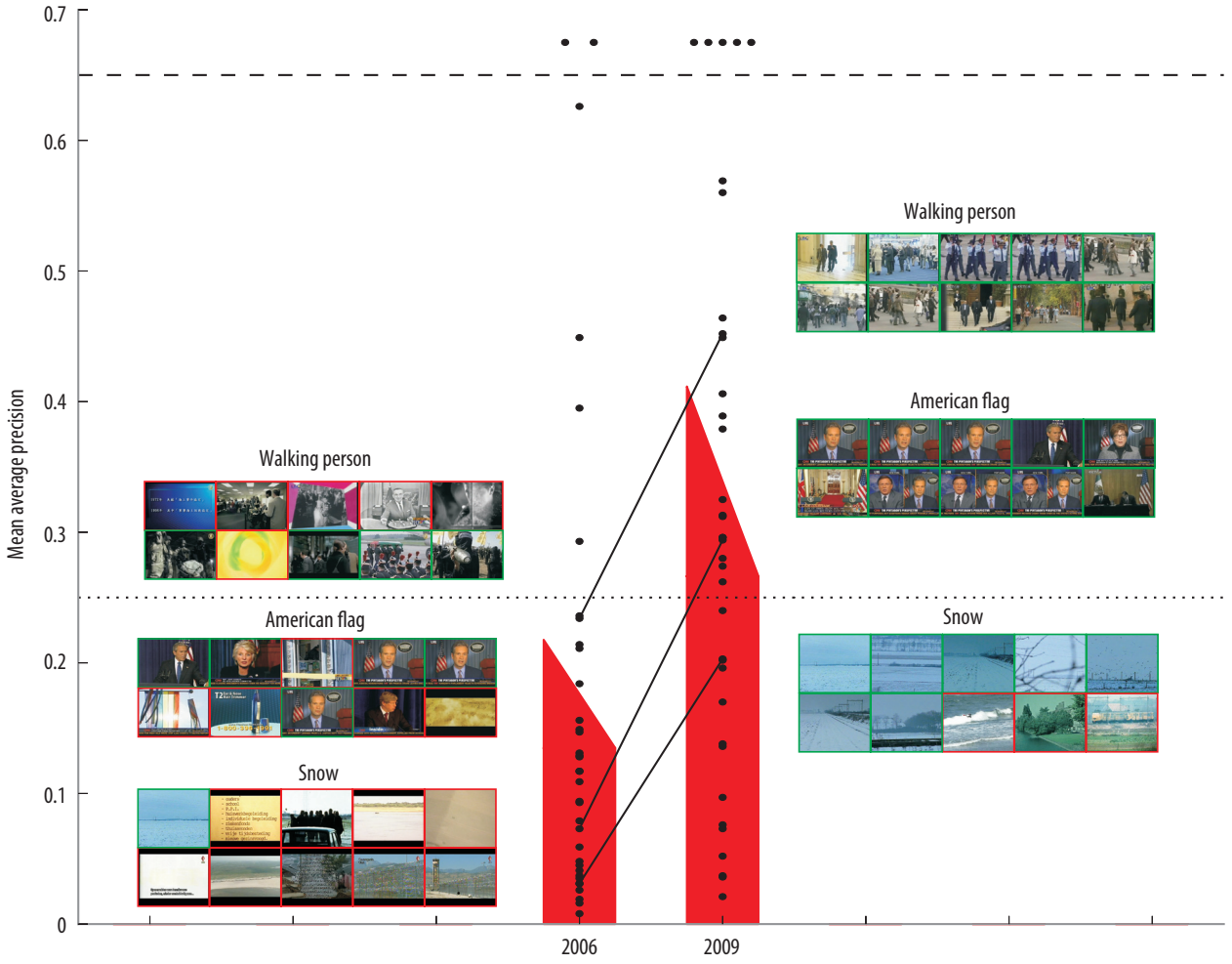

Figure 2. Visual-search progress as evaluated on 36 concept detectors (•) derived from broadcast video data using state-of-the-art search engines from 2006 and 2009. The figure highlights performance for three typical concepts. The top of the skewed bar indicates the maximum average performance by training on similar examples, and the bottom indicates the minimum performance when training on a data set of completely different origin. A mean average precision score of 0.25 (dotted line) is generally accepted to be sufficient for interactive search. The horizontal dashed line represents Google's text-search performance. Contrary to belief in the community, progress in visual search is substantial and visual-concept search is quickly maturing in robustness for real-world usage of any concept.

the situation where the training set data was visually similar to the testing set, and that where the training set data visually differed from the set used for testing (www.mediamill.nl/progress).

As Figure 2 shows, search engine performance doubled in just three years. For learned concepts, detection rates degenerated when applied to data of a different origin yet still doubled in three years. Thus, contrary to the widespread belief that visual-search progress is incremental and detectors generalize poorly, our experiment shows that progress is substantial on both counts.

Progress may be greater than expected, but it doesn't imply that the general problem of visual search is solved. Our experiment used only 36 concepts, whereas broad visual search would require thousands of detectors approaching the vocabulary of a common user. Nevertheless, we believe that broad categorization, as the first step toward semantic interpretation of images, is within reach.

Cees G.M. Snoek *is a Veni research fellow in computer science at the University of Amsterdam. Contact him at cgmsnoek@uva.nl.*

Arnold W.M. Smeulders *is a professor of computer science at the University of Amsterdam. Contact him at arnoldsmeulders@uva.nl.*

Editor: Naren Ramakrishnan, Dept. of Computer Science, Virginia Tech, Blacksburg, VA; naren@cs.vt.edu