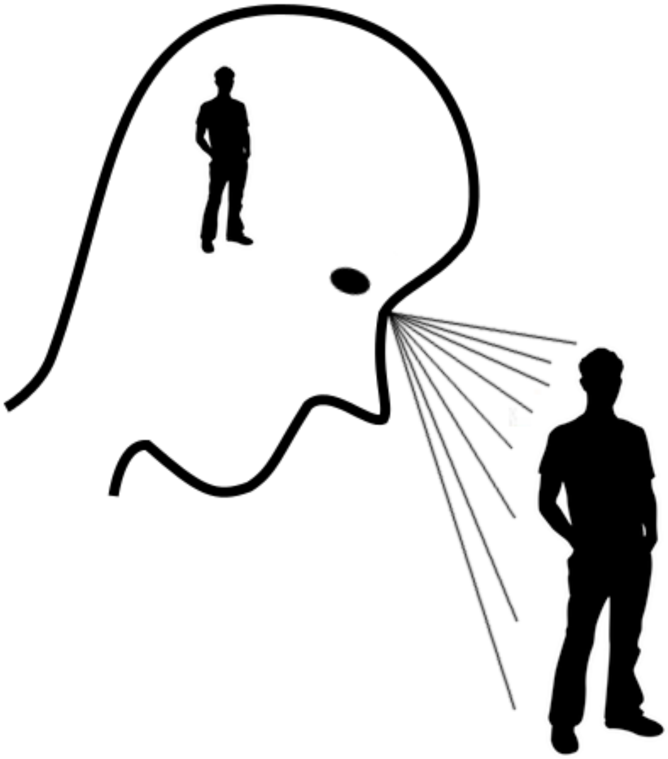
Unlocking the power of video with intelligent computing

How software which mimics the human brain can better recognize objects in video streams.



A White Paper by Sighthound, Inc.

Introduction

We are about to emerge from the dark ages of computer video. Until now computers have been blind - they have depended entirely on humans to tag or label video to do anything useful with them. As a result, we can use only a fraction of the content recorded by millions of video cameras every day. Soon, the ability to identify objects in video will become a baseline requirement for any video application or website. It will enable a wide range of new applications in entertainment, advertising and video search.

The security industry pioneered video analytics, aiming to increase the efficiency of people watching video monitors. These efforts led, however, to a disparate collection of often closed and expensive systems that are unreliable in the real world. This paper reviews the limitations of conventional video analytics, and describes a fundamentally different approach offered by a product called Sighthound Video.

Video analytics today

Consumer-grade solutions

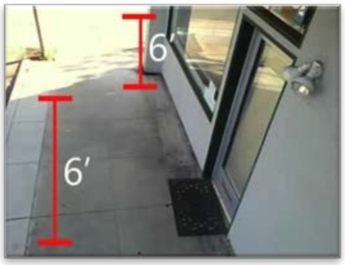
Off-the-shelf security applications available in consumer channels use motion detection to determine if moving objects are present in video. Motion detection analyzes how many pixels have changed between frames. When enough pixels change, video is recorded and/or the customer is notified of a "motion event." Unfortunately, pixels can also change with clouds passing overhead, leaves swaying in the wind, or flickering lights. More advanced systems let customers ignore selected regions in the video, or set sensitivity levels. Nevertheless, motion detection often generates an unacceptable number of false alerts. Customers are flooded with notifications and recordings of empty scenes. The typical consumer story goes something like this: There's a theft from a house in the neighborhood. The homeowner installs video cameras for peace of mind. The software that comes with the system is dauntingly complicated to set up and use. If the consumer does manage to set up the product, the feature that allows alerts to be sent to a smartphone triggers hundreds of false alarms. By day three the homeowner has turned off the alerts.

Motion detection can be triggered by light changes, reflections or swaying shadows.

Professional-grade solutions

Systems that identify objects require more expensive equipment, and expert technicians to calibrate and tune each camera. As a result, sales are limited to qualified corporate and government customers, through vertical sales channels. Today's analytics systems use a two-step process to recognize people in videos. First, pixel changes are analyzed to isolate moving objects from the background. Boxes are drawn around these objects. Second, "object recognition" algorithms are applied to classify the moving objects as people, vehicles or other objects.

Conventional object recognition starts with the assumption that vertical boxes are humans, and horizontal boxes are not. Rules about the boxes are then added to improve accuracy, starting with size. Boxes that are much shorter or taller than humans are filtered out. This requires that a technician calibrate the 3D scene to a 2D video image by entering camera height, lengths of objects at varying places in the screen, etc. In addition, the motion of the box may be analyzed to assess speed, consistency of movement, etc. Some systems attempt deeper analysis of the visual content inside the box. This may sound straightforward until one considers the tremendous variation of shapes in the real world: people carrying umbrellas covering their heads, pushing strollers or carrying large items, "split" when walking behind objects, etc. To a rules-based recognition system, these cases represent radical departures from the canonical form of a person. Expanding the rules to accommodate this variation unfortunately increases the number of non-human forms that also meet those criteria.



Drawing lines on a screen to calibrate the ground plane and perspective.

To compensate for these shortcomings, conventional systems use cameras with higher resolution, which in turn increases data storage and creates network bandwidth bottlenecks. As a result, analytics is being pushed to the camera, requiring embedded processors that push up cost. And more consulting and technician support is required to install, calibrate and maintain these high-end systems.

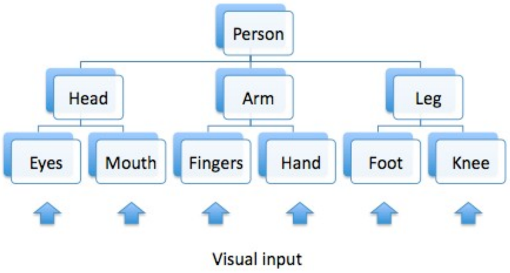
Sighthound Video: a new approach

Rather than relying on expensive equipment and expert human tuning, Sighthound Video uses smarter algorithms that work with *less* expensive equipment, out of the box. The algorithms are based on a theory of human intelligence developed by Jeff Hawkins, the inventor of the PalmPilot. This theory, called hierarchical temporal memory (HTM), describes how humans instinctively compensate for dizzying variation that confounds rules-based programming.

How the brain recognizes objects

According to HTM theory, the brain is a hierarchical learning network. Each node at the bottom of the network sees a small stream of input data. Over time it starts to categorize familiar patterns. In the vision example, it might realize that it tends to see simple lines, edges or curves. It then passes that information up to the node above it, which in turn starts to categorize patterns of patterns received from the nodes below it. One combination of lines and edges might be a hand, another

combinations of curves a mouth, etc. That node then passes that information to the node above it, to generate a more complex form (e.g., a pattern of eyes and mouth might be a head). The node at the top of the hierarchy only receives the input from the nodes below it, perhaps telling it that there is a head and arm and leg visible, and labels that pattern as a "person."



Simplified conceptual image of how the brain recognizes a person.

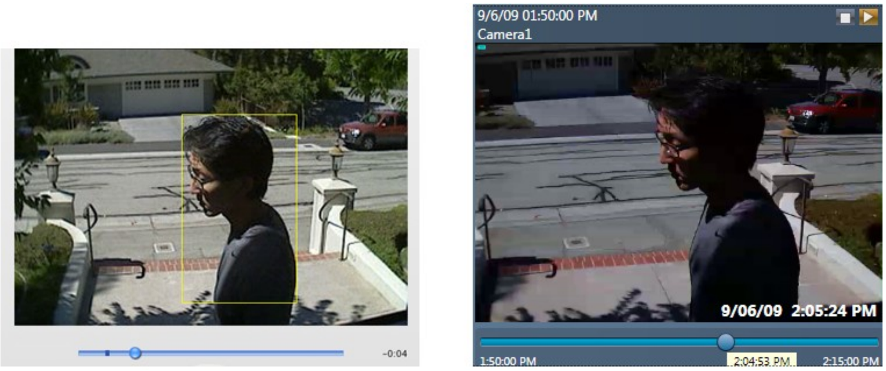
It does not directly analyze the millions of pixels in the image, and is not confused by nearly infinite variations of raw data. If a brain sees enough body components, it will conclude it is looking at a human form, even if large parts are hidden or unclear. As a result, learning the objects takes considerable time, but when new variations of the objects are seen, recognition is fast and effective.

Sighthound Video vs. conventional programming: applying brain theory

Inspired by the human intelligence model, Sighthound Video abandons the approach of rules based on box shape and movement. Instead, a learning network was trained with video clips of humans. The network was taught that these video clips represented people, and trained to learn that examples of vehicles and animals were not people. As a result, Sighthound Video has demonstrated a high degree of tolerance to real-world challenges relative to conventional methods.

In good conditions, Sighthound Video can generate recognition accuracy of 90-95%. Sighthound defines "good" conditions as scenes where most of the people are clearly visible and walking upright. More challenging scenes include cases where a large part of the scene is under dark shadows, or where many people walk behind large trees or other objects, or at angles where people aren't fully in frame as they pass.

The screen shot on the left below is from a scene where Sighthound Video generated 92% accuracy over the course of an afternoon. The screen shot on the right was taken from a state-of-the-art professional system that, in a side-by-side comparison, generated a 75% accuracy figure over the same time period. The screen shot shows an actual failure case seen relatively frequently on the professional system in this experiment: persons who are less than 10 or 20 feet from the camera.



Sighthound Video successfully detects a person while professional grade system fails to detect the same person.

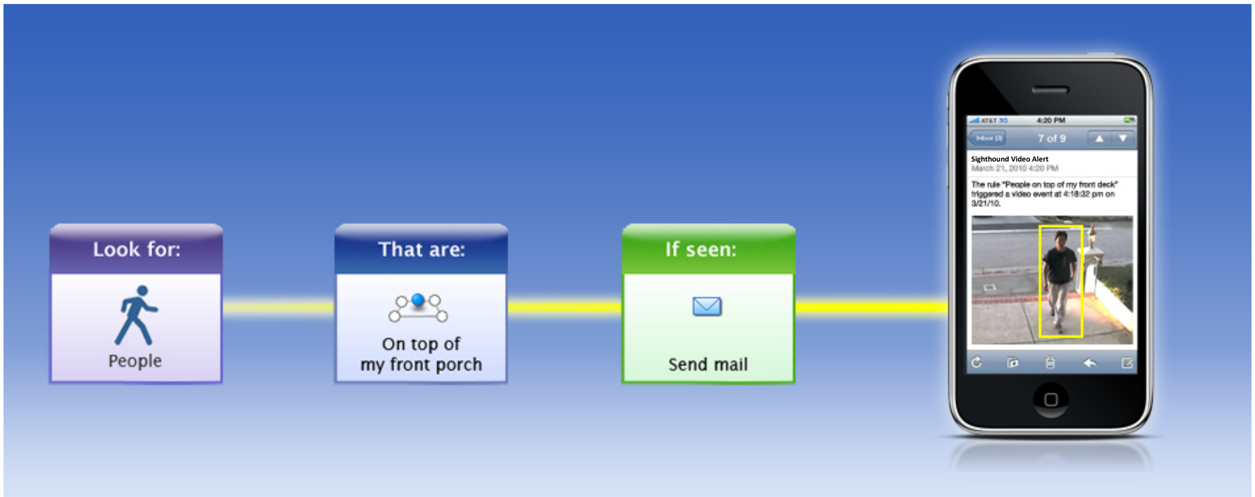
The failure probably occurs because only part of the body is visible. Raw accuracy data is not meaningful out of context, however, since accuracy varies dramatically between and within scenes. The 75% accuracy figure can be increased considerably by mounting an array of cameras so that people are detected in the distance before approaching, and the entire body can be seen. This is likely what a reseller installing a professional system would recommend. On the other hand, the accuracy would drop significantly if a single camera were mounted to detect people outside a door. This is an important use case for the small business and residential customers that Sighthound Video targets.

Solving the right problem, in the most elegant way

Even the best raw technology must target an appropriate problem. Video analytics has been marketed as failsafe protection of critical assets in large deployments. Unfortunately, not only did conventional systems fail to live up to the accuracy claims, but also the solutions they targeted were inappropriate in the first place. Even with a 99% accurate system, false positives can be crippling in the real world. At 99% accuracy, 2000 objects appearing in a day would trigger 20 false alarms.

Sighthound Video, on the other hand, is optimized for rapidly refining searches and reviewing video. The goal is to distill an unwieldy number of video clips into a manageable amount, and to quickly refine searches when looking for a specific event. This usage pattern is more closely aligned to the reality of accuracy that is below 100%.

In this context, user interface plays a role that is as critical as accuracy. Sighthound Video was designed to refine searches in order to reduce significantly the amount of video to scan.



Sighthound Video Alert
March 21, 2010 4:20 PM

The rule "People on top of my front deck" triggered a video event at 4:18:32 pm on 3/21/10.

Finally, eliminating the need for technicians to calibrate and maintain the system drives down installation and total cost of ownership. It also enhances overall system flexibility by allowing a customer to easily move a camera for a temporary or specific need.

Sighthound sees this approach as a template for solving information overload problems. The key to dealing with too much information is to find relevant information with intelligent computing, and to provide a streamlined interface to find what you want quickly. It becomes another big data problem, but with more CPU processing power and storage than either text or speech.

Future potential

Since Sighthound Video is based on a learning system, it has the potential for far more sophisticated recognition problems than conventional video analytics. Box-shape analysis has attempted to address more complex recognition problems. Relative speed and location of the boxes can indicate running or loitering, for example. There are fundamental limitations, however, when applying rules-based systems to other problems.

For example, the box-shape rules described above would have no meaning in the context of recognizing faces or other types of objects. Sighthound Video is based on a training approach that can be expanded to different types of objects, or different types of human actions (such as walking, climbing or throwing). The same methods can be applied to problems ranging from optical character recognition to facial recognition. Moreover, end users will be able to teach the system to recognize new objects through a process of labeling video clips without any programming required.

Summary

The rules-based approaches of conventional AI have failed to live up to their expectations, and have inherent limitations in generalizing to broader sets of problems. Intelligent computing, on the other hand, promises to solve a host of real-world problems that have challenged conventional programming for decades.

Sighthound Video leverages intelligent technology to powerful effect. Intelligent computing, combined with thoughtful interface design, will create simple solutions to complex problems.