Brian Dipert and Amit Shoham

Eye, Robot

Embedded vision, the next big thing in digital signal processing.

Full-featured, power-stingy, and low-cost processors, image sensors, and other system building blocks are poised to make embedded vision the next digital signal processing success story. Fueled by high-volume consumer electronics applications, expanding to a host of other markets, and brought to life by a worldwide alliance of suppliers and system developers, the technology is sure to fulfill its promise in the near future.

Digital signal processing has been the forefront of BDTI's (Berkeley Design Technology's) focus since the company's founding in 1991, as an extension of the founders' work in developing digital signal processing design tools, methodologies, and architectures at the University of California at Berkeley. Over the intervening 20 years, BDTI has worked closely and frequently with Gene Frantz and other DSP visionaries at Texas Instruments, along with their counterparts at other companies. And abundant historical evidence over that two-decade timespan validates the longstanding symbiotic relationship between digital signal processing and the consumer electronics market.

Digital Object Identifier 10.1109/MSSC.2012.2193077
Date of publication: 13 June 2012

1943-0582/12/$31.00©2012IEEE

Consider, for example, wireless networking. Previously implemented in a proprietary fashion for many years in niche applications, it finally achieved the necessary price points for consumer electronics consideration in the second half of the 1990s. Earlier in the decade, the Global System for Mobile Communications (GSM) standard brought digital signal processing to wireless WAN communications, both for voice and data traffic; the contending code division multiple access (CDMA) approach followed shortly thereafter. Subsequent digital signal processing case studies in consumer electronics include digital audio processing, both in the living room and the shirt pocket, and still and video image capture using digital bits instead of silver halide flakes.

Embedded vision, the next likely digital signal processing success story, refers to machines that understand their environment through visual means. It leverages (for example) the sensors and SoCs in the previously mentioned digital cameras in taking processing to the next level: interpreting meaning from and responding to the information in the captured frames. We use the term *embedded vision* to refer to any image sensor–inclusive, microprocessor-based system that isn't a general-purpose computer (see Figure 1). It could be applied to a smart phone, a tablet computer, a surveillance system, an earthbound or flight-capable robot, a vehicle containing a 360-degree suite of cameras, or a medical diagnostic device. Or it could be a wired or wirelessly tethered user interface peripheral; Microsoft's Kinect for the Xbox 360 game console, perhaps the best-known example in this last category, sold 8 million units in its first two months on the market.

Embedded vision can alert you to a child struggling in a swimming pool and to an intruder attempting to enter your residence or business. It can warn you of impending hazards on the roadway around you and even prevent you from executing lane-change, acceleration, and other maneuvers that could be hazardous to yourself and others. It can equip a military drone or other robot with electronic "eyes" that enable limited or even fully autonomous operation. It can assist a human physician in diagnosing a patient's illness. It can uniquely identify a face in front of the image sensor and subsequently initiate a variety of actions: automatically logging into a user account, for example, or displaying relevant news and other information. It can interpret gestures and even discern a person's emotional state. And, in conjunction with GPS, compass, accelerometer, gyro, and other sensors, it can present a data-augmented representation of the scene encompassed by the image sensor's field of view. These and innumerable other applications provide ripe opportunities for participants in all areas of the embedded vision ecosystem, computer vision veterans and new entrants alike.

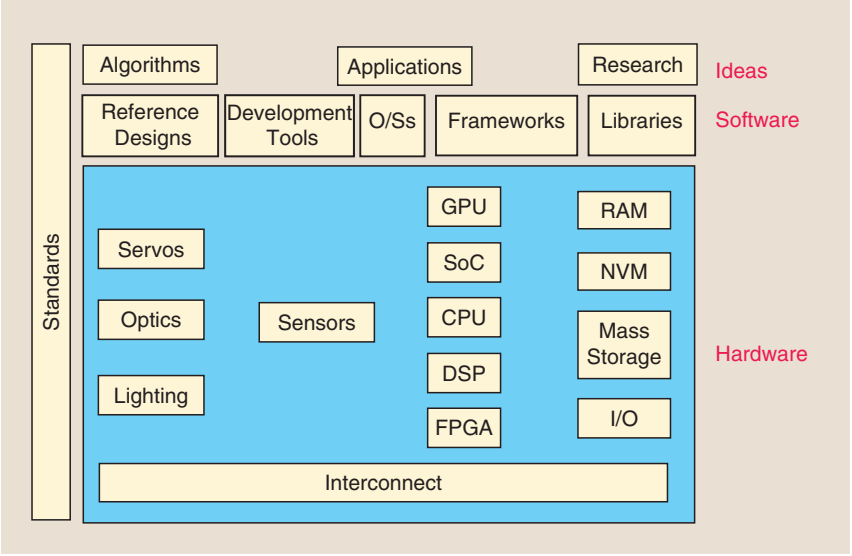

FIGURE 1: The embedded vision ecosystem spans hardware, semiconductor, and software component suppliers; subsystem developers; systems integrators; and end users, along with the fundamental academic research that provides ongoing implementation breakthroughs. A unified worldwide alliance will enable the ecosystem to thrive to the richest possible degree.

Embedded vision can alert you to a child struggling in a swimming pool and to an intruder attempting to enter your residence or business.

Embedded Vision Leverages Digital Signal Processing

A typical embedded vision processing flow consists of three main functional stages (see Figure 2):

- image acquisition and optimization
- converting pixels into objects
- analysis of—and reasoning about—these objects.

Each stage in the overall process leverages a substantial amount of digital signal processing. Image acquisition and optimization, for example, usually encompass at least a subset of the following functions:

- noise reduction
- image stabilization

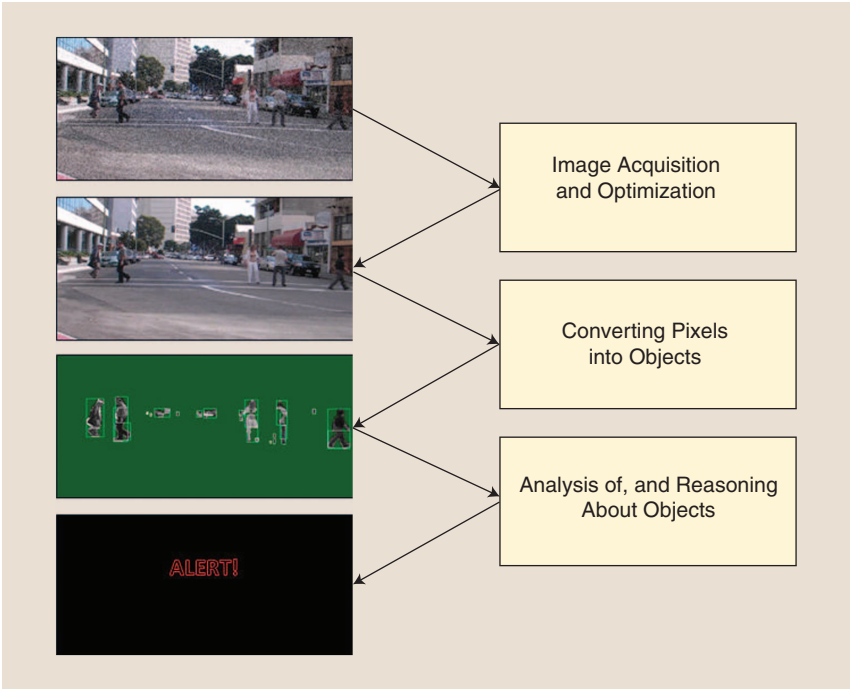

FIGURE 2: The embedded vision application pipeline spans three primary stages, whose respective functions extensively harness digital signal processing algorithms.

- lighting compensation
- lens distortion correction
- color space conversion.

Keep in mind that optimization may mean something completely different in embedded vision applications versus traditional image capture scenarios. Processing for embedded vision may result in an image that isn't particularly pleasing to the human eye but simplifies the subsequent steps of locating features of interest and identifying objects based on those features. For example, consistency in illumination of objects from one frame to the next may be very important while aesthetically pleasing illumination is irrelevant.

When considering the process of converting pixels into objects, keep in mind that specific applications within this broad category tend to use unique sets of functions, and consequently there may be little common ground among the applications. They all heavily leverage digital signal processing techniques implemented in hardware and/or software, however. Some functions implement algorithms that extract useful information about pixels and groups of pixels. They might, for example, evaluate the motion at each pixel position, determine if a pixel represents an edge or corner, determine if a pixel is statistically interesting (i.e., outside of normal bounds) in some way, determine depth at a pixel position from a stereo image, and so on. Functions in this subcategory include:

- image filtering: 2-D finite impulse response (FIR), 3-D FIR, gradient filters, Haar filters, temporal filters, and so on
- edge, corner, and feature detection
- image threshold determination
- histograms and other statistics
- optical flow
- background subtraction (including the adaptive and multimodal variants)
- morphological functions, such as erosion and dilation.

Other common functions group pixels into "objects" that can be correlated with real-world equivalents. A function might determine, for example, that a set of edge pixels forms a line, a useful capability for detecting lane markings on a road. Other functions might determine that a set of red pixels represents a stop sign; that a cluster of pixels is part of an in-motion, to-be-tracked person or vehicle; or that a suite of pixels makes up a hand or a face. Functions in this subcategory include:

- connected component labeling
- contour tracing
- clustering
- the Hough transform (for line and circle-or-ellipse detection)
- curve, surface, and mesh-fitting algorithms.

While many applications distinctly implement the function subcategories mentioned above, in other applications these subcategories are blended together. For example, an application might leverage a cascade of Haar filters for face detection or object detection functions. In general, the task of converting pixels into objects requires "touching" every pixel of every frame. Each function must therefore process a significant number of pixels (more than 120 million per second, for example, in the case of a 60-fps 1080p video stream), thereby leading to very high computational and memory bandwidth requirements.

From Objectification to Identification and Reaction

Functions that involve analyzing and reasoning about objects often don't need to "touch" any pixel data; instead, they rely solely on the object data supplied by the functions described earlier. Sometimes, however, the identification and reaction category overlaps with the preceding ones. In order to reason about objects, we may need additional information about them. The functions mentioned earlier may have identified and objectified a cluster of interesting pixels, for example, that we now want to correlate with objects analyzed in previous frames. As such, we might need to answer questions such as "What color is this object?" or "Where is the object's center?" While we may have already extracted

relevant information via the image capture and objectification functions, we may still need additional pixel-level information to obtain all of the necessary information.

Regardless, key distinctions between this category and the prior ones include the fact that these particular functions don't touch every pixel, don't process the same data (the same image region, for example) of each frame, and, depending on the application, may not be invoked at all for some frames. For example, if we haven't detected any objects in the current frame, we're not going to compute the centers of any objects (conversely, if we've detected lots of objects, we may need to decide which ones are most interesting by computing multiple objects' centers, among other things). The narrowed focus in this category translates into much lower computational demands, along with lower I/O and memory bandwidth needs, unless the application uses extremely sophisticated algorithms or analyzes vast numbers of objects.

At a high level, functions in this category may implement:

- object movement tracking
- object classification (e.g., determining if an object is a person, a vehicle, or something else)
- object recognition (e.g., determining whose face has been detected)
- obstacle detection (deciding whether an object lies in our path)
- behavior recognition (e.g., identifying a hand-gesture command)
- behavior classification (e.g., understanding whether a person is walking, sitting, standing, or lying down)
- prediction (e.g., given an object's trajectory and/or history, anticipating where will it be in two seconds and what it will be doing at that time.

Algorithms used to accomplish these functions, many of which leverage various digital signal processing techniques, are extremely diverse and include:

- predictive filters (Kalman filters are popular)
- statistical analysis and modeling (including hidden Markov models, histograms, and norms)
- back-projection, correlation, and computation of various error metrics
- finite state machines, heuristics, and rule-based decision trees
- database searches
- neural networks.

Embedded vision technology has the potential to enable electronic products to be more intelligent and responsive so that they are more valuable to users.

Making Embedded Vision's Potential a Reality

Embedded vision technology has the potential to enable electronic products to be more intelligent and responsive so that they are more valuable to users. It can let electronic equipment companies both create valuable new products and add helpful features to existing products. And it can provide significant new markets for hardware, semiconductor, and software manufacturers. The Embedded Vision Alliance (EVA), a unified, worldwide organization of embedded vision ecosystem representatives, will transform this potential into reality in a rich, rapid, and efficient manner.

The EVA has developed a robust Web presence (www.embedded-vision.com), freely accessible to all and including (among other things) online seminars, technical articles, and a multisubject discussion forum staffed by a diverse group of technology experts. The EVA's Web site exemplifies the organization's primary mission of inspiring and empowering embedded vision innovation and adoption through cultivation of awareness and education. Other aspects of the EVA's charter include:

- the incorporation and commercialization of technology breakthroughs originating in universities and research laboratories around the world
- the hosting of comprehensive education facilities that will enable new players in the embedded vision application space to rapidly ramp up their expertise
- the proliferation of hardware and software reference designs and other development aids that will enable system implementers to develop products that optimally meet unique application needs.

For more information, please visit www.embedded-vision.com and contact the EVA at info@embedded-vision.com or 1-510-451-1800.

About the Authors

Brian Dipert is a senior analyst with Berkeley Design Technology, Inc. (BDTI) and the editor in chief of the Embedded Vision Alliance. Prior to joining BDTI and the EVA, he spent more than 14 years as a senior technical editor at *EDN Magazine*; before joining *EDN*, he spent three years at Magnavox Electronic Systems Company and eight years at Intel Corporation. He holds a bachelor's degree in electrical engineering from Purdue University.

Amit Shoham is an distinguished engineer with BDTI. At BDTI, he provides technical leadership on a wide variety of projects focused on evaluating, improving, and creating processors, tools, algorithms, software, and system designs for embedded digital signal processing applications. He holds a bachelor's degree in computer systems engineering and a master's degree in electrical engineering, both from Stanford University.

SSC