Event-Based Vision enables the next revolution in visual perception for machines

Author: C. Posch and S. Lavizzari

Description of the innovation:
Event-based Vision (EBV) is poised to take over from the frame-based approach used by traditional film, digital and mobile phone cameras in many machine-vision applications. Event-based solutions will make machine vision more accessible and more effective, enabling a world in which production lines run more quickly, humans and machines work together more safely, autonomous vehicles are accident-free, and drones intuitively avoid collision at high speeds.

The mode of operation of state-of-the-art image sensors is useful and efficient for exactly one thing: photography, i.e. for taking an image of a still scene. Exposing an array of pixels for a defined amount of time to the light coming from such a scene is an adequate procedure for capturing its visual content. Such an image is a snapshot taken at one point in time and contains zero dynamic information. Nonetheless, this method of acquiring visual information is also used in practically all machine vision systems for capturing and understanding dynamic scenes. This approach is seemingly supported by the way movies are made for human observers. The observation that visual motion appears smooth and continuous if viewed above a certain frame rate is, however, more related to characteristics of the human eye and brain than to the quality of the acquisition and encoding of the visual information as a series of still images. As soon as change or motion is involved, which is the case for almost all machine vision applications, the paradigm of visual frame acquisition becomes fundamentally flawed. If a camera observes a dynamic scene, no matter where you set your frame rate to, it will always be wrong! As different parts of a scene
usually have different dynamic contents, a single sampling rate governing the exposure of all pixels in an imaging array will naturally fail to yield adequate acquisition of these different scene dynamics present at the same time.

An "ideal" image sensor samples parts of the scene that contain fast motion and changes at high sampling rates and slow changing parts at slow rates, all at the same time - with the sampling rate going to zero if nothing changes. Obviously, this will not work using one common single sampling rate, the "frame rate", for all pixels of a sensor. Conversely, one wants to have as many sampling rates as there are pixel in the sensor - and let each pixel's sampling rate adapt to the part of the scene it sees.

To achieve this requires putting each individual pixel in control of adapting its own sampling rate to the visual input it receives. This is done by introducing into each pixel a circuit that reacts to relative changes of the amount of incident light, so defining the individual pixel's sampling points in time (Fig. 1). As a consequence, the entire image data sampling process is no longer governed by a fixed (artificial) timing source (the frame clock) but by the signal to be sampled itself, or more precisely by the variations over time of the signal in the amplitude domain (Fig. 2). The output generated by such a camera is no longer a sequence of images but a time-continuous stream of individual pixel data, generated and transmitted conditionally, based on what is happening in the scene.

Following this paradigm, we have developed an image sensor containing an array of autonomously operating pixels that combine an asynchronous level-crossing detector with a separate exposure measurement circuit (Fig. 1). Each exposure measurement by an individual pixel is triggered by a level-crossing event. Inspired by biology, every pixel in these sensors optimizes its own sampling depending on the visual information it sees. In case of rapid changes, the pixel samples at a high rate. On the contrary, if nothing happens, the pixel stops acquiring redundant data and goes idle until things start to happen again in its field of view. Hence each pixel independently samples its illuminance upon detection of a change of a certain magnitude in this same luminance, thus re-measuring its new light level after it has changed. The result of the exposure measurement (i.e. the new gray level) is asynchronously output off the sensor together with the pixel's coordinates in the sensor array (Fig. 3). As a result, image information is not acquired and transmitted frame-wise but continuously, and conditionally only from parts of the scene where there is new visual information. Or in other words, only information that is relevant - because unknown - is acquired, transmitted, stored and eventually processed by machine vision algorithms. This way, both the acquisition of highly redundant and useless data by over-sampling static or slow parts of the scene, and the under-sampling of fast scene dynamics due to fixed frame rates, can be eliminated.

Pixel acquisition and readout times of milliseconds to microseconds are achieved, resulting in temporal resolutions equivalent to conventional sensors running at tens to hundreds of thousands of frames per second. Now, for the first time, the strict temporal resolution vs. data rate tradeoff that limits all frame-based vision acquisition can be overcome. As the temporal resolution of the image data sampling process is no longer governed by a fixed clock driving all pixels, the data volume of the sensor output,
independently of the temporal resolution available for the acquisition at the single pixel, is only
dependent on the dynamic contents of the visual scene. Visual data acquisition simultaneously becomes
fast and sparse, leading to ultra-high-speed acquisition combined with reduced power consumption,
transmission bandwidth and memory requirements.
The advantage of treating dynamic visual information this way does not end at the sensing stage. In order
to fully unlock the potential of these event-based vision sensors, also the paradigms of vision processing
need to be fundamentally rethought. First of all, the notion of a frame at the basis of vision processing is
to be abandoned altogether. As the sensors encode visual dynamics into highly resolved spatio-temporal
patterns of events, representing the relevant features of scene dynamics (such as moving object contours,
trajectories, velocity, etc.), processing algorithms now work on continuous time events and features
instead of on discrete static images. The mathematics that describe these features in space and time are
simple and elegant and yield highly efficient algorithms and computational rules that allow for real-time
operation of sensory-processing systems while minimizing demand on computing power.
The materialization of our research effort led to the launch of the most advanced event-based reference
system: ONBOARD (Fig. 4). It integrates the new 3rd generation VGA sensor camera module with MIPI
CSI interface, into a powerful reference vision system ARM-based Quad Core platform. It provides
comprehensive connectivity including Ethernet, USB, HDMI, Wi-Fi, operating under a Linux OS. The
embedded system runs dedicated computer vision software. Currently, it offers a tracking algorithm to
detect motion, segment data into groups of spatio-temporal events and track over time (taking 2 out of 4
available cores). The application layer comprises area monitoring, high-speed counting, vibration
measurement and real-time inspection, which makes ONBOARD the perfect fit for the new artificial vision
capabilities demanded by the 4th industrial revolution.

Technical details and advantages of the innovation:
Among the advantages of Event-Based Vision are: fast sensing at low cost, ability to operate in poor
and/or variable lighting, low computational, memory and communications overhead and low energy use.
PROPHESEE’s event-based vision sensor (Fig. 5) is based on a VGA (640x480) array of fully autonomous
pixels containing event-based change detection and pulse-width-modulation (PWM) imaging circuitry.
Exposure measurements are initiated and carried out locally by the individual pixel that has detected a
change of brightness in its field-of-view. Pixels do not rely on external timing signals and independently
and asynchronously request access to an (asynchronous arbitrated) output channel when they have new
grayscale values to communicate. Pixels that are not stimulated visually do not produce output. The visual
information acquired from the scene, temporal contrast and grayscale data, are communicated in the
form of asynchronous address-events (AER), with the grayscale values being encoded in inter-event
intervals. The pixel-autonomous and massively parallel operation ideally results in lossless video
compression through complete temporal redundancy suppression at the pixel level. Compression factors
depend on scene activity and go beyond 1000 for static scenes. Due to the time-based encoding of the illumination information, very high dynamic range—intra-scene DR of 143 dB static and 125 dB at 30 fps equivalent temporal resolution—is achieved. A novel time-domain correlated double sampling (TCDS) method yields array FPN of

**Relevance and application possibilities of the described innovation for the machine vision industry:**
The 4th industrial revolution, driven by automatization of processes and tasks, interactions between autonomous machines and between humans and machines, not forgetting scalability, is demanding for new artificial vision capabilities.

Area monitoring: tracking algorithm can now run on embedded computers, directly at the edge. The algorithms capture all kind of motion in the scene: operators movement, machines moving, AGV/forklifts, local lights changes and even flies flying in front of the camera. Parameters allows to configure the tracking to discard some movements, based on velocity, amount of time during which an object remains static, size, region of non-interest (RONI).

Realtime counting: drastically improve productivity by counting and measuring objects moving across a field of view at rates of thousands of pieces per second, in real time and with a compact, cost-efficient system. Event-based vision strongly reduces the need for more expensive ultra high-speed, matrix or line scan cameras, reducing the cost and complexity of your set up. Objects are counted as they pass through the field of view, triggering each pixel independently as the object goes by. By only recording the pixels independently triggered by changes, capturing the essential information the system requires and no more. This new approach allows for unprecedented counting speed that can reach thousands of counts per second. Moving objects are counted at time resolutions of 10’s of microseconds and are not subject to motion blur. Because each pixel is independently triggered by motion, objects can never move more than one pixel between two acquisitions. This means event-based algorithms can track objects smoothly, even at very high speeds.

Vibration monitoring: the speed of acquisition and processing allow to measure the vibration frequency of tools and machines in real time and from a distance in a noninvasive way, opening new possibilities in tool monitoring and predictive maintenance, e.g. by visually detecting a changing in the frequency and amplitude of oscillation of an engine in can be inferred the health status and eventually performing technical maintenance before the engine breaks or the production tool is forced to unscheduled shutdown.

Viewing and realtime inspection of welding, flames and high intensity processes: thanks to the extremely high dynamic range, event-based vision can visualize and monitor processes in high dynamic range conditions, such as welding, that are difficult to image with conventional camera systems. In addition, high speed tracking can be directly applied to high dynamic range processes: visualization and measurements for arc welding, visualization and measurements for laser welding and visualization of metal arc spraying, including both melt regions and surroundings.
Automated machines and robots: fast visual feedback loops for robotics systems ensure safer co-working between robots and humans. Complex multi-object, real-time, detection-recognition-tracking, improves the capabilities of the automated guided vehicles (AGV): fast detection enabling faster automatic emergency braking. Lower costs and computational overhead will enable the use of more cameras for increased redundancy and safety. Kinetic monitoring, coupled with real-time energy-efficient SLAM algorithms, enables new opportunity in the packaging and robot domain.

Video:
https://youtu.be/pTdBBIrzrhSk

Images:
39574_prophesee_fig1_atis.jpg
39574_prophesee_fig2_sparsity.jpg
39574_prophesee_fig3_xyt.jpg
39574_prophesee_fig4_onboard-reference-system.jpg
39574_prophesee_fig5_sensor.jpg