A Battery Powered Vision Sensor for Forensic Evidence Gathering

Y. Zou
M. Lecca
M. Gottardi
Fondazione Bruno Kessler
Trento, Italy
(zou,lecca,gottardi)@fbk.eu

G. Urlini
STMicroelectronics
Agrate Brianza, Italy
giulio.urlini@st.com

N. Vretos
L. Gymnopoulos
P. Daras
Centre for Research & Technology
Hellas (CERTH) - ITI
Thessaloniki, Greece
(vretos,lazg,daras)@iti.gr

ABSTRACT

We describe a novel battery-powered vision sensor developed to support surveillance and crime prevention activities of the Law Enforcement Agencies (LEA) in isolated or peripheral areas not equipped with energy grid. The sensor consists of a low-power, always-on vision chip interfaced with a processor executing visual tasks on demand. The chip continuously inspects the imaged scene in search for events potentially related to criminal acts. When an event is detected, the chip wakes-up the processor, normally in idle state, and starts delivering images to it together with information on the region containing the event. The processor re-works the received data in order to confirm, to recognize the detected action and in case to send the alert to the LEA. The sensor has been developed within a H2020 EU project and has been successfully tested in real-life scenarios.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

Vision sensors, low-power camera, motion detection.

ACM Reference Format:
Y. Zou, M. Lecca, M. Gottardi, G. Urlini, N. Vretos, L. Gymnopoulos, and P. Daras. 2019. A Battery Powered Vision Sensor for Forensic Evidence Gathering. In *ICDSC 2019 13th International Conference on Distributed Smart Cameras, September 09–10, 2019, Trento, IT*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Video-surveillance of large areas plays an increasing role for supporting the activities of crime discovery and crime prevention conducted by the Law Enforcement Agencies (LEAs) worldwide. Nevertheless, covert evidence gathering has not seen major changes in decades. LEAs are even today using conventional, man-power based techniques such as interviews and searches to gather forensic evidence. As sophistication of both criminals and their crimes are increasing, investigators must improve the tools available for gathering a body of compelling evidence of a suspect's involvement in a crime. Concealed surveillance devices have been instrumental in this direction, providing irrefutable evidence that can play an important part in bringing criminals to justice. However, current video surveillance systems are usually bulky and complicated, and rely on complex, expensive infrastructure to supply power, bandwidth and illumination. In addition, the integrity of the acquired digital evidence plays a predominant role in the digital process of forensic investigation. Proper chain of custody must include information on how the evidence was collected, transported, analysed, preserved, and handled. It must document where, when and how the digital evidence was discovered, collected, handled with, when and who came in contact with the evidence and whether it is altered in any way. If a link is missing in this chain, it could be deemed compromised and may be rejected by the court.

Recent years have seen significant advances in the surveillance industry, in both hardware and software, but these were often targeted to other markets and applications. Sensor technology has evolved, resulting in smaller modules with significantly higher resolution and image quality. At the same time platforms that host the sensors have also improved, embedding more powerful processing units, enabling more complex but also more energy-demanding operations. The imaging community is currently focusing on cameras for mobile phones, where the figures of merit are resolution, image quality, and extremely low profile. Power consumption, while an important parameter, is often a secondary aspect. A mobile phone with its camera on would consume its entire power supply in less than two hours. Industrial surveillance cameras are even more power hungry, typically requiring 10W for their operation, even without night illumination. Vision algorithms for scene understanding are currently not embedded in such systems, rather demanded to external clouds that achieve high-level scene interpretation but also require extremely high processing power. Although these new hardware and software technologies represent important advances, they are inappropriate for video-surveillance tasks conducted in isolated, peripheral areas, that are often chosen by criminals to perpetrate their activities. These places include nature parks, wooded areas and beaches, extra-urban areas that for their location, environmental constraints or costs cannot be equipped with electric grid.

ICDSC 2919, September 09–10, 2019, Trento, IT

https://doi.org/10.1145/1122445.1122456

In this context, low-power, battery-operated systems represent an adequate, respectful of the ambient solution.

In this paper we present a novel, battery-powered, wireless sensor for evidence gathering, able to operate without infrastructure. The sensor consists of an always-on VGA CMOS imager embedding an event detection algorithm and of a processor, that is woken up by the vision chip to classify every detected event and in case of a crime, to send an alert to LEAs. The combination of built-in intelligence with low power consumption makes the proposed device a true breakthrough for combating crime.

2 SENSOR ARCHITECTURE

As shown in Figure 1, the sensor architecture consists of a custom vision chip, acquiring images, running a low-level algorithm to detect events aimed at triggering an external processing platform executing high-level algorithms on data that are delivered by the vision chip, a communication module, a communication network, and the backend. In this Section, we will briefly describe the architectures of the two main system components, the vision chip and the processor, and the low-level and high-level algorithms embedded in the image sensor and in the processor respectively.

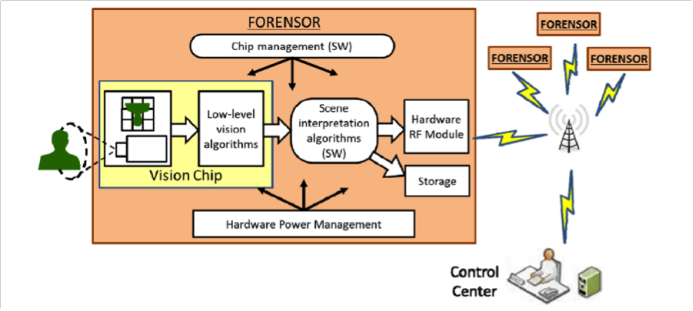


Figure 1: Block diagram of the vision sensor architecture.

2.1 Vision Chip

The custom low-power CMOS vision chip continuously analyzes the scene, delivering data only in case of alert conditions. The 640 x 480 pixel rolling-shutter imager, shown in Figure 2, embeds an image processing layer based on 160 Processors executing robust motion detection through a double-threshold pixel-wise background subtraction on a sub-sampled image of 160 x 120 pixels. The voltage V_{pix} of each pixel is compared with two thresholds (Vmin, Vmax), acting as low-pass filters, approaching the negative and the positive peaks of the signal respectively. Figure 3 shows how the low-level algorithm is applied on each pixel. Vmax and Vmin follow the maximum and minimum values of the pixel respectively, defining a safe zone (Vmin < Vpix < Vmax) outside which the pixel is said Hot-Pixel (HP), detecting a potential alert situation. At the end of the comparison, the two thresholds are updated and stored into the on-chip SRAM, to be retrieved next frame.

The equations, regulating the background subtraction algorithm

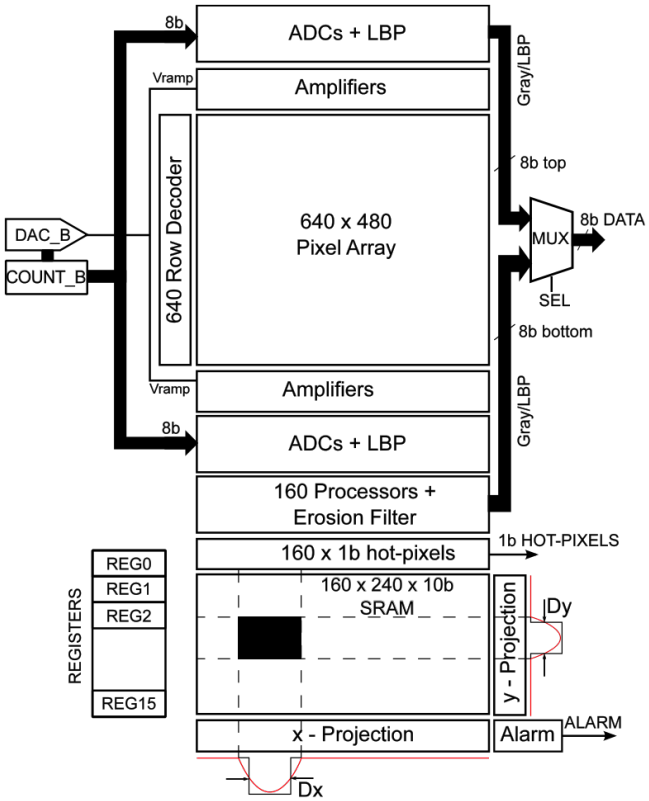


Figure 2: Block diagram of the low-power VGA vision chip.

for each pixel, are the following:

$$Vpix > Vmax \Rightarrow Vmax = Vmax + \Delta H; \quad (1)$$

$$Vpix \leq Vmax \Rightarrow Vmax = Vmax - \Delta L; \quad (2)$$

$$Vpix < Vmin \Rightarrow Vmin = Vmin - \Delta H; \quad (3)$$

$$Vpix \geq Vmin \Rightarrow Vmin = Vmin + \Delta L; \quad (4)$$

with ΔH and ΔL (with $\Delta H > \Delta L$) programmable values, shared among all pixels of the array. If the pixel value (Vpix) is larger or lower than ΔHOT, above VH or below VL, a Hot-Pixel (HP) is asserted:

$$Vpix > Vmax + \Delta HOT \ or \ Vpix < Vmin - \Delta HOT. \quad (5)$$

At the end of each frame, a 120 x 160 pixels motion bitmap is generated (where white pixels are HPs) and de-noised by a bank of 160 programmable Erosion Filters operating on a (3 x 3) pixel kernel. The HPs contribute to the generation of two xy-projection vectors which are low-pass filtered and binarized against two user-defined thresholds, Dx and Dy. ALARM is generated only when the HPs of the xy-projection vectors form a region with a pre-defined size and aspect ratio (e.g. black rectangle in Figure 1).

The vision chip also embeds a digital processing block executing Local Binary Pattern coding [6] computed over a 3 x 3 pixel kernel. It is integrated near the ADCs (Analog to Digital Converters) and data can be delivered sharing the 8 bit output DATA bus. In the application described here LBPs are not used, however they have been integrated in the chip for further possible uses, such as hand

gesture or face recognition [1]. Figure 4 shows how the sensor processes images aimed at detecting events. The main parameters of the vision chip as reported in Tab. 1. The vision sensor prototype is shown in Figure 5 b). It adopts a low-power FPGA to control the chip and to implement the interface with the processor.

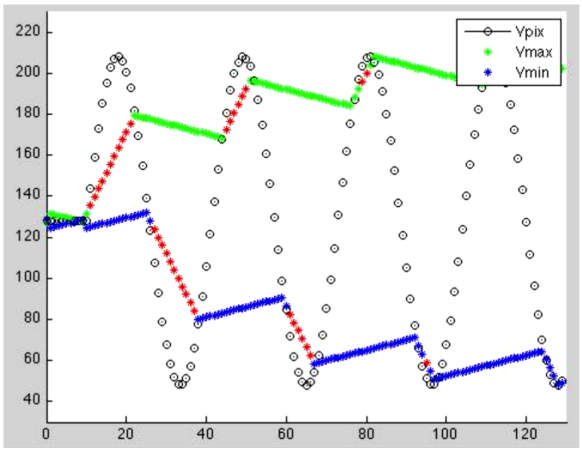


Figure 3: Simulation of the pixel-wise, double-threshold background subtraction algorithm over 130 frames. The red-dotted lines represent the frames with Hot-Pixels.

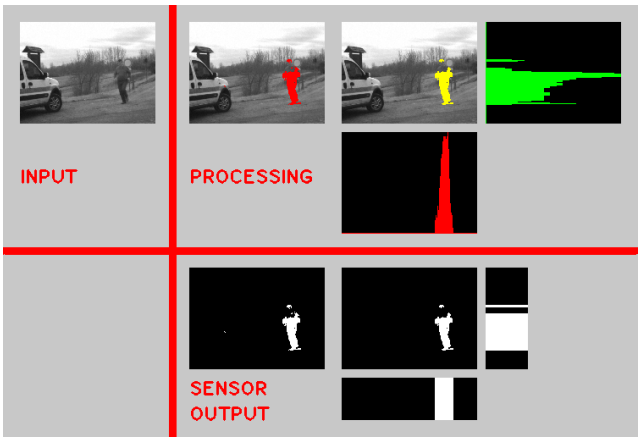


Figure 4: An example of image processing performed by the vision chip. Input: acquired image; Processing: Hot-Pixels before (red) and after (yellow) the erosion filter, and the xy-projections of the Hot-Pixels. Sensor Output: 120x160 HP bitmap after de-noising and related binarized projections.

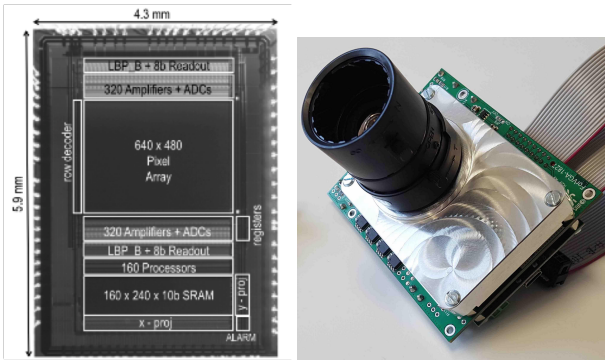
2.2 The processor

The processor is the local computing unit responsible for three activities:

(1) the acquisition of the images from the vision chip and the interpretation of the video stream;

Table 1: Main vision chip characteristics

Parameters	Value
Technology	110 nm CMOS
Pixel size	4 μm x 4 μm
Fill factor	49%
Dynamic range	53 dB
Supply voltage	3.3V / 1.2V
Power consumption (8 fps)	344 μW (motion)
	1350 μW (imaging)
Features extraction	Temporal change + xy-projection
ALARM generation	xy-projection



a) b)

Figure 5: a) Microphotograph of the vision chip; b) module hosting the vision chip with the interface for the processor.

(2) the activation of the communication to the gateway about the events, state of the device, any accessory alert;
(3) the power management of the system, i.e. the control and switch of the sensor and processor states from the idle to the running mode;
(4) the execution of the high-level vision algorithm, that takes in input the event detected by the vision chip (specifically, the corresponding gray-level frames and their Hot-Pixel maps with the xy-projections), classifies the action detected as an event by the vision chip and if this is labeled as "potential crime" send an alert to the LEAs.

The processor, called SecSoC, is a System on Chip (SoC) device based on a heterogeneous multi-core DSP for video processing applications. The device is characterized by four cores to be programmed with low level applications for the data processing of the video images acquired and for the selection and transmission of given results. The generic purpose cores are supported in the computation by several hardware-accelerated functions. A high-level picture of the SecSoc chip internal architecture is depicted in Figure 6, where the main box in the centre represents the System on Chip and its elements. The objects surrounding it are the possible connected peripherals. The four DSP cores can be identified in the left lower box of the picture where four R4DSP cores are depicted. The core is a STRED4 processor, a proprietary ST core, version four

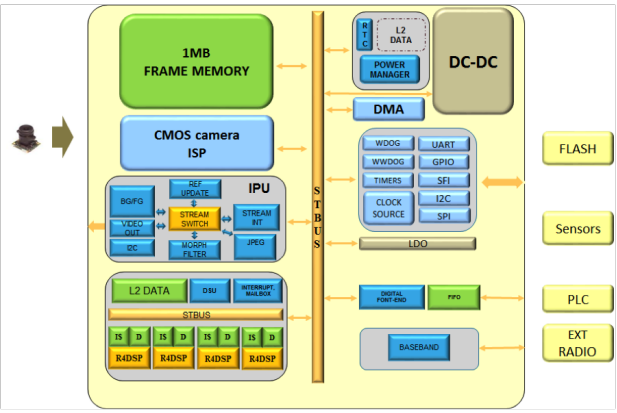


Figure 6: Architecture of the System on Chip SecSoC.

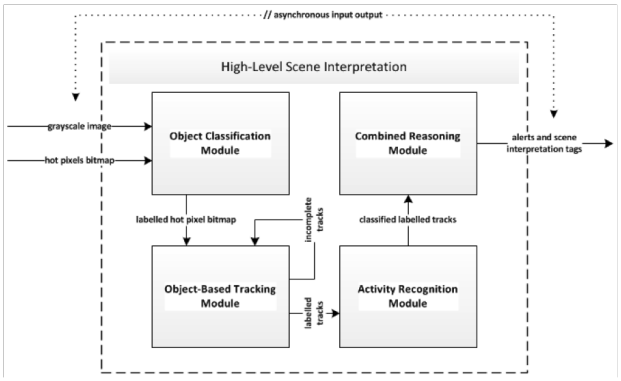


Figure 7: Algorithmic flow within the High-level Scene Interpretation blocks.

of the family of STRED processors. They have been designed for an ultra-low power consumption processing, with a devoted instruction set. Each core is equipped with a fast local memory for data and instructions. A global bank of RAM memory of 1Mb is accessible by all the cores and the IPU sub-module. The stream of uncompressed raw data is acquired by the video sensors (on the left side of the box). The data is processed by the Image Signal Processing (ISP) block that provides a set of basic accelerated functions for the processing of the raw data. The data can be directed to the Image Processing Unit (IPU) block for further processing, as described in paragraph ISP and IPU Hardware accelerated functions. The configuration and control of these blocks is left to the programs implemented on the four cores.

2.3 High level Algorithm

The High-Level Algorithm (HLA) embedded on the SecSoc is responsible for the classification of the event detected by the vision sensor. This classification includes the recognition of the objects/people involved in the event, the motion analysis and the scene understanding, i.e. recognition of the event as a licit action or as a certain crime.

The algorithmic flow of HLA (see Fig. 7) includes these steps:

- Given a trigger/ alert from the low-level block, the high-level block is activated;
- The high-level block receives Regions of Interest (ROIs, described by the Hot-Pixel bitmap and, if required, a corresponding gray-scale image) from the low-level block;
- The high-level block performs an object classification of the provided ROIs, producing a labelled Hot-Pixel bitmap;
- The labelled Hot-Pixels are tracked along consecutive frames;
- Event classification is performed on the labelled tracks;
- The classified labelled tracks are finally combined to produce alerts and scene interpretation tags.

The object detection algorithm consists of 2 steps:

(1) Computation of a feature vector describing the 2D silhouette of the ROI in the Hot-Pixel map;
(2) Classification of the object descriptors by a linear SVM model.

To compute the feature descriptor for the pre-processed binary input image, the HLA component initially computes the centroid of the ROI pixels:

$$c_y = \frac{1}{\sum_{x,y} \beta_{x,y}} \sum_{x,y} \beta_{x,y} y \tag{6}$$

$$c_x = \frac{1}{\sum_{x,y} \beta_{x,y}} \sum_{x,y} \beta_{x,y} x \tag{7}$$

where $\beta_{x,y} \in \{0,1\}$ is the binary pixel value of the Hot-Pixel map at position (x,y) and (c_x, c_y) is the centroid coordinates.

Therefore, HLA component defines a number of bins B in which it partitions the angular space around the foreground centroid (see Figure 8 for an example). Subsequently, it forms a descriptor f of size $1\times$ B counting the total number of non-zero pixels contained in each region of the partitioned angular space. To compute in which region of the partitioned angular space each pixel (x,y) of the Hot-Pixel map belongs to, HLA calculates the pixel Four-Quadrant Inverse tangent with respect to a Cartesian coordinate system centred about the point (c_x, c_y) with the X and Y axis aligned in parallel to the image axis. More precisely:

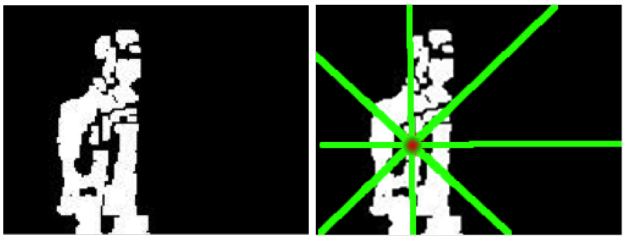
$$\phi_{x,y} = \arctan \frac{(y - c_y)}{x - c_x} \in [0, 2\pi) \tag{8}$$

Then, each pixel (x,y) with $\beta_{x,y} = 1$ contributes a vote to the i-th component of $f(i \in \{1,2,...,B\})$ with i given by the following formula:

$$i = \text{floor}(\frac{\phi_{x,y}}{2\pi} B) + 1 \tag{9}$$

The feature f is then L-1 normalized and passed down to the linear SVM for inference. In conclusion, this descriptor is translation invariant and essentially captures the shape of the silhouette of the foreground object. The time complexity of this algorithm is linear with the number of pixels in the image, and thus of low computational complexity.

The evaluation of the discriminative capabilities of the object descriptions has been carried out on 1188 Hot-Pixel binary images of humans and cars captured by the vision chip and such that the number of images depicting car being equals the number of images depicting human (i.e. a 50% distribution of images along the classes). Moreover, for each class, the 40% of images have been used for training while the remaining 60% for testing. The classification accuracy of our proposed algorithm reaches 94% demonstrating



a) b)

Figure 8: a) Hot-Pixel map of a moving person processed by morphological filters; b) Angular space partitioning around the foreground centroid.

its suitability for our use cases. For porting the object description algorithm onto SecCoC, its routines were firstly rewritten in plain C programming language and tested on a PC. Then the code was modified to be compatible with the hardware platform.
The object classification algorithm is divided in three steps:

- **Feature extraction**, which is performed on the connected components of the Hot-Pixel map, computed as in [8].
- **Region Description**: the algorithm first calculates the object position and then its centroid. We do not calculate the centroid of different clusters but only the centroid of the main cluster, i.e. the detected object. Next, the angle between the line segment, defined by each non-zero pixel and the centroid and the x-axis is computed. The descriptor is then normalised based on the sum of the non-binary pixels and finally normalised with the use of the maximum and minimum values of each bin that have been computed during the training procedure.
- **Object Classification**: a linear SVM has been implemented on SecSoc. The weights and the bias that have been extracted from the training procedure are stored in the code segment of SecSoc software as they were hard coded in a C header file. The training of the model is done offline in MATLAB using data recorded by the sensor during the pilots. Apart from its acceptable evaluation performance, the linear SVM model has been chosen for the classification task since it can be nicely fit into SecSoc memory, in contrast to our previous attempts of fitting a Convolutional Neural Network (CNN) in the embedded system.

The developed software was uploaded on SecSoc and has been tested in real conditions, producing reliable results. It is important to mention that the features extracted by the real system differ slightly from those computed by the high-level software version, because each hardware architecture has a different floating-point precision. The difference in the result of each operation is negligible, as it is less than 0.01%.

3 EXPERIMENTAL RESULTS

The system has been tested in real-life scenarios: identifying illegal pedestrian intrusion into reserved areas, detecting illegal access in reserved roads and helping combat international drug trafficking into Europe.

3.1 Evaluation of On-Chip Event Detection

In order to validate the performance of the event detection algorithm, the vision chip was tested separately. Fig. 9 shows a snapshot extracted from a 62s video taken with the sensor at 8 fps. The two outputs (VGA grey-scale image and QQVGA Hot-Pixel bitmap) are delivered by the vision chip, which monitors a boat approaching the coast in a windy scenario. The resulting motion bitmap (top right) is compared with a second bitmap (bottom right) obtained processing the grey-scale image with frame-difference. The latter is a technique largely used in other vision chips [2] [3] [4], being straightforward to implement on-chip. However, its main drawback is that it is very sensitive to noise. In fact, while in our case the boat is clearly detected, in frame-difference it cannot be even distinguished from the background. Table 2 reports a comparative analysis of the main characteristics of the proposed sensor versus other similar motion-based sensors. Although its Figure Of Merit (FOM), i.e. the energy needed by the chip in one frame divided by the number of pixels, is worse than the above chips, the main advantage is its larger reliability which has been demonstrated experimentally in noisy scenarios, as shown in Figure 9.

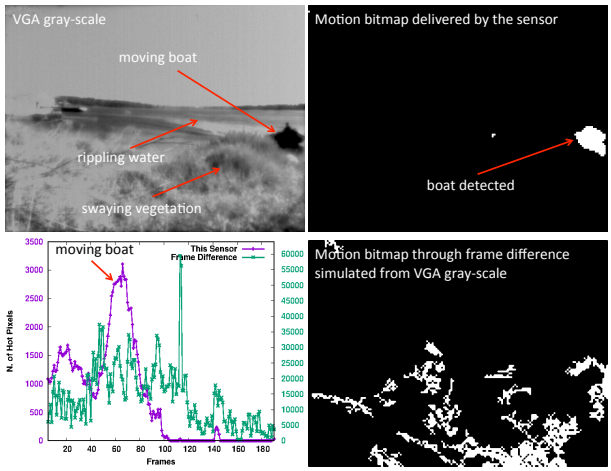


Figure 9: Vision chip outputs in a real scenario. The graph compares the detection behavior of the proposed sensor vs. frame difference.

3.2 Evaluation of HLA

The evaluation of the HLA has been conducted on two different public datasets along with our proper pilot testing. The chosen datasets where the KTH dataset [5] and the MuHaVi dataset [7]. Finally, in order to choose the parameters of the system, we ran a grid based approach. The KTH dataset contains 600 videos from 25 people performing 6 different actions. The actions included in the dataset are: Walking, Jogging, Running, Boxing, Hand Clapping and Hand Waving. For evaluation we use the standard procedure proposed in [5]. The total recognition accuracy achieved by our method is 88.7%. On the other hand, the MuHaVi is a rich dataset containing 17 action classes in total that were performed by 7 people and each action was captured from 8 cameras. Such actions are "Walk and then turn back", "Run and Stop", "Punch", "Pull Heavy Object" and

Table 2: Comparison of the main sensor parameters against similar motion-based vision sensors

Parameter	[2]	[3]	[4]	This work
Technology	0.35μm CMOS 1P4M	0.18μm CMOS 1P4M	0.13μm CMOS 8M1P	0.11μm CMOS 1P4M
Pixel array	128 x 64	256 x 256	128 x 128	640 x 480
Motion Resolution	128 x 64	128 x 128	48 x 168	160 x 120
Motion Detection	Frame difference	Frame difference	Frame difference	Double-threshold + erosion
Alert generation	pixel count	pixel count	pixel count	xy-projections
Features extraction	Temporal change + Binary contrast	Temporal change + HOG	Temporal change	Temporal change + xy-projections + Local Binary Pattern coding (LBP)
Pixel pitch	26 μm x 26 μm	5.9 μm x 5.9 μm	6.4 μm x 6.4 μm	4 μm x 4 μm
Fill Factor	20%	30%	38%	49%
Dynamic Range	-	54dB	38.5dB	53dB
Supply Voltage	1.3V/0.8V	1.3V/0.8V	1.2V/0.6V	3.3V/1.2V
Power consumption	100 μW@50fps	51.06 μW@15fps (imaging) 3.31 μW@15fps (motion)	29 μW@19fps (imaging) 1.1 μW@30fps (motion)	1.35 mW@8fps (imaging) 344 μW@8fps (motion)
FOM (imaging) W/pixel · frame	244 pW	51.94 pW	152 pW	549 pW

Table 3: Results of the event detection Algorithm on different Datasets. See text for more details.

Dataset	KTH	MuHaVi	Real Life Pilots
Recognition Rate	88.7%	92.8%	82.4%

others. While MuHaVi is a multi-view dataset, we use a single view for each action and more precisely the view which is perpendicular to the action. The evaluation strategy is set to leave-one-actor-out. The total recognition accuracy of the proposed method for all the 17 classes is 92.8%. Finally, In the FORENSOR Pilots the actions were "Car passing by", "boat arriving at the coast", "boat leaving the coast", "Pedestrian walking", "Pedestrian Standing" as advised by the different Law Enforcement Agencies (LEAs) involved in the pilots. We have achieved an 82.4% accuracy rate in this last dataset as previously annotated by the LEAs. All LEAs involved in the piloting were commenting on the usefulness of such apparatus and its power of performance. Moreover, the low rate of false alarms was something that impressed the different LEAs representatives. The recognition accuracy rates obtained on the three test sets are summarized in Table 3.

4 CONCLUSION

The paper reports on a vision sensor targeted to forensic evidence gathering. The adopted event-driven paradigm allows computing resources to be used only upon request, largely improving the energy-efficiency of the whole system. This is even more true in outdoor scenarios where noise and uncontrolled lighting condition might heavily affect the vision system performance. In our case, the vision chip works as the master of the system, detecting events in the scene and asking the processor, normally in idle mode, to execute high-level algorithms only when necessary. By so doing, while the vision chip is always-on, the duty-cycle of the processor and of the wireless communication is drastically reduced. A robust vision chip-embedded event detection algorithm is therefore extremely important to minimize the false positives that otherwise would turn on the system uselessly with a large waste of power.

5 ACKNOWLEDGMENTS

This work was funded by the EU H2020-FCT-2014 FORENSOR (FOREnsic evidence gathering autonomous seNSOR) Project, Grant n. 653355 (https://cordis.europa.eu/project/rcn/194854).

REFERENCES

[1] A. Berkovich, M. Lecca, L. Gasparini, P. Abshire, and M. Gottardi. 2015. A 30 μW 30 fps 110 x 110 Pixels Vision Sensor Embedding Local Binary Patterns. *IEEE Journal of Solid State Circuits* 4, 9 (Sept 2015), 2138–2148.

[2] M. Gottardi, N. Massari, and S. A. Jawed. 2009. A 100μW 128x64 Pixels Contrast-Based Asynchronous Binary Sensor for Sensor Network Applications. *IEEE Journal of Solid State Circuits* 44, 5 (May 2009), 1582–1592.

[3] J. Choi and S. Park and J. Cho and E. Yoon. 2013. A 3.4μW CMOS Image Sensor with Embedded Feature-Extraction Algorithm for Motion-Triggered Object-of-Interest Imaging. In *IEEE International Solid-State Circuits Conference Digest of Technical Papers*. 478–479.

[4] G. Kim, M. Barangi, F. Zhiyoong, N. Pinckney, B. Suyoung, D. Blaauw, and D. Sylvester. 2013. A 467nW CMOS Visual Motion Sensor with Temporal Averaging and Pixel Aggregation. In *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2013 IEEE International*. 480–481.

[5] Ivan Laptev, Barbara Caputo, et al. 2004. Recognizing human actions: a local SVM approach. In *null*. IEEE, 32–36.

[6] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence* 24, 7 (2002), 971–987.

[7] Sanchit Singh, Sergio A Velastin, and Hossein Ragheb. 2010. MuHaVi: A multicamera human action video dataset for the evaluation of action recognition methods. In *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*. IEEE, 48–55.

[8] Kenji Suzuki, Isao Horiba, and Noboru Sugie. 2003. Linear-time connected-component labeling based on sequential local operations. *Computer Vision and Image Understanding* 89, 1 (2003), 1–23.