8th INTERNATIONAL WORKSHOP ON ON-BOARD PAYLOAD DATA COMPRESSION (OBPDC 2022)

28-30 SEPTEMBER 2022

The VIDEO project: Enabling RGB video processing and compression on-board satellites

Yubal Barrios, Romén Neris, Raúl Guerra, Sebastián López and Roberto Sarmiento

Institute for Applied Microelectronics (IUMA), University of Las Palmas de Gran Canaria (ULPGC)
Campus Universitario de Tafira S/N, 35017 Las Palmas de Gran Canaria, Spain
Email: {ybarrios, rneris, rguerra, seblopez, roberto}@iuma.ulpgc.es

INTRODUCTION

Video acquisition systems are gaining interest for space missions because they send a continuous flux of data which is especially relevant in Earth monitoring and deep-space exploration missions. This information is also useful for target detection and specially for tracking purposes, monitoring in real-time the desired area to provide a quick response under specific events.

Nonetheless, video sensors collect a high amount of data, which cannot be neither stored nor sent in raw format to ground due to the limited downlink bandwidth, preventing near real-time capabilities [1]. These constraints will become more stringent during the next years, since next-generation sensors will increase both the pixel resolution and the scene size. Additionally, the video post-processing performed on ground prevents from an immediate response to certain situations, such as natural phenomena, surveillance or security purposes [2], [3].

For these reasons, the development and implementation of solutions for on-board video processing and compression is recommended. On-board target detection and tracking use the high volume of information provided by video sensors in both the temporal and the spatial domains to provide a quick response under unexpected events, while on-board video compression takes advantage of the high temporal correlation between consecutive frames in a video sequence to efficiently reduce the data volume.

In this work, the H2020-funded Video Imaging Demonstrator for Earth Observation (VIDEO) project is presented, which is focused on the development of a next-generation instrument for Earth observation [4]. The VIDEO instrument will have the capability to perform high-resolution video monitoring on an extremely wide scene, with the purpose of recognizing and tracking objects. The processing finishes by compressing the useful data (i.e., the video segment where the target has been detected) to minimize the downloaded information to ground.

The rest of the paper is structured as follows. First, an overview of the VIDEO project goals is provided to put this work into context. Then, the video processing chain, in which the ULPGC/IUMA is the main responsible, is described in detail, including both the target detection and the video compression stages. Then, the status of the video processing chain is provided, including some preliminary results. Finally, the main conclusions achieved in this project at the moment of presenting this work are summarized.

THE VIDEO PROJECT

The VIDEO project will be a new type of instrument designed to be used with next generation of Artificial Intelligence processing capacity on-board. Due to its specific and innovative technologies and architecture adapted to new generation of data processing, the VIDEO instrument will be the pathfinder of the next instrument generation for Earth Observation.

The VIDEO project proposes a set of breakthrough technologies for instruments for Earth Observation. Indeed, the VIDEO project is the future of the small and compact instrument with extra wide field of view. Based on TAS' exclusive patent combining freeform mirrors in a smart optical compact combination and latest advances in material development, the VIDEO instrument will have the capability to collect high resolution images as well as video monitoring on an extremely wide scene. The Field of View for this type of imager instruments will be about 10 times higher in surface than the equivalent instrument volume without freeform disruptive solution.

On top of that, the idea of the VIDEO project instrument is to combine the latest technologies in terms of Additive Manufacturing for low Coefficient of Thermal Expansion (CTE) AlSix material development, in order to have same material for structure and mirrors. AlSix is a material specifically developed in order to adapt its CTE to the desired value thanks to the proportion of Si added in the Aluminium matrix ("x" refers to the proportion of Si in the material). Thanks to that, the instrument will have an extremely stable and homothetic behaviour for an optical point of view.

The consortium includes 6 entities from 3 different European countries (Belgium, France and Spain), combining the skills of academics and SMEs, and three large industrial companies, among which the project coordinator, Thales Alenia Space. The VIDEO project will include the latest advances in Europe technologies for telescope structure and mirrors as well as latest innovative solution for video detection, acquisition, treatment and compression.

In this latter point is in which the University of Las Palmas de Gran Canaria (ULPGC), the only academic partner, is the main responsible of object detection (mainly ships, the use case of the VIDEO project), tracking and video compression. This implies the selection of the suitable algorithms for each video processing stage and their implementation on hardware embarked on-board satellites.

VIDEO PROCESSING CHAIN OVERVIEW

The video chain of the VIDEO instrument shall include adaptive algorithms able to change the output parameters of the image such as video rate, compression ratio, compression with or without losses... These algorithms will be aware of the type of image being acquired, for example to increase the compression rate when flying over the ocean or reducing the video rate (or even stopping it) when capturing clouds. In the same way, they will adapt without ground intervention to changes on the satellite environment.

Thanks to the wide field of view thought for the VIDEO instrument (e.g., 24 km per 15 km square FoV at 500km altitude), permanent travelling without slow motion can be consider for the spacecraft, which is here denoted as *detection* or *image mode*. This permanent travelling detection mode will allow typically few seconds video or imaging with very downgraded signal and resolution (due to high-speed swath on ground). But thanks to the smart algorithms implemented on-board, it will allow to identify motion or shape that will automatically trigger the *identification* mode with permanent video monitorization. The *identification* or *video* mode will be possible with classical satellite attitude compensation (quasi-infinite slow motion) in order to allow about 2 minutes video mode on the same field of view.

During the identification mode, a GSD <1m and 10 Hz acquisition is considered, enough to allow to the video processing algorithms to detect and analyse object motion on the scene. During the video identification mode, the algorithm will decide to perform multi-window acquisition if relevant, depending on objects detected and identified in the whole scene. Data to be downloaded will be chosen among the relevant windows analyzed to minimize the size of the downloaded data, which is finally compressed before sending it to ground. Both the detection and the identification modes are graphically reflected in Fig. 1.

In order to accomplish the on-board real-time processing demands of the VIDEO project, a video processing chain has been fully developed using High-Level Synthesis (HLS) techniques, in order to quickly obtain a functional prototype of the whole system.

The target detection stage implements the MobileNetv1Lite convolutional network, demonstrating high accuracy ship detection results from Remote Sensing scenes [5], which is the use case defined for the VIDEO project. This CNN, though it is not the best one in terms of performance (measured in GFLOPs), provides a trade-off between its precision and the number of trainable parameters, process that is exercised on ground.

On the other hand, the video compression stage implements an extended version of the CCSDS 123.0-B-2 near-lossless compression standard, originally thought to compress hyperspectral images, capable of handling RGB video sequences by introducing some pre-processing stages that also contribute to increase the compression performance. The compressor is also able to apply different errors taking into account Regions of Interest (ROI) in a frame by indicating the coordinates of that region [6]. In the case of the VIDEO project, the object detected is preserved with a high level of detail (lossless compression), while the rest of the frame (mainly sea) is compressed with a high error. In addition, panchromatic video compression is supported, if needed, by directly reusing a CCSDS 123.0-B-2 processing core and substituting the spectral by the temporal domain [7].

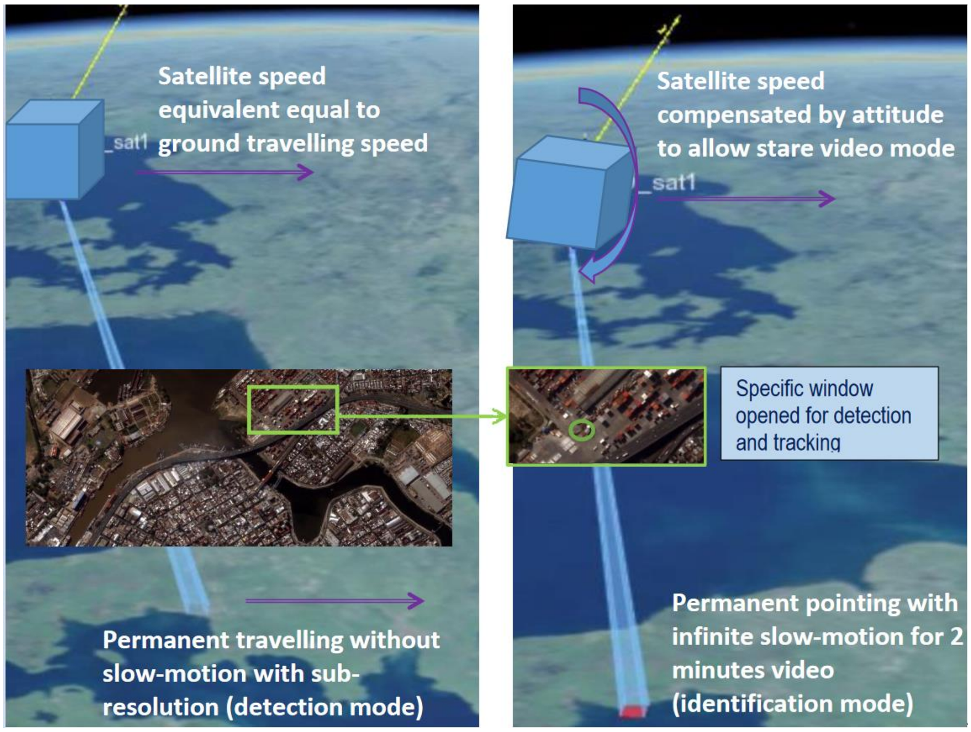
Satellite speed equivalent equal to ground travelling speed
_sat1
Permanent travelling without slow-motion with sub-resolution (detection mode)
Satellite speed compensated by attitude to allow stare video mode
_sat1
Specific window opened for detection and tracking
Permanent pointing with infinite slow-motion for 2 minutes video (identification mode)

Fig. 1. VIDEO instrument acquisition modes

The VIDEO instrument will integrate an RGB sensor, which acquires the scene in a line-by-line way (i.e., a whisk broom scanner), so the information is provided in Band-Interleaved-by-Line (BIL). Therefore, both the target detection and the video compression stages will work in BIL order, processing a complete line of samples in the spatial domain before starting the next band, avoiding any extra intermediate stage to reorder the image pixels before feeding the video processing chain and thus allowing real-time capabilities.

VIDEO PROCESSING CHAIN STATUS

System design

The RGB sensor developed by Pyxalis in this project will provide a considerably large image in terms of spatial resolution (48Mpixels), so it is not feasible to store the entire image in memory and start feeding the CNN with the complete image so that the CNN can process it in one shot. Besides, the image encoding method used by the sensor is BIL, so waiting for the whole image to arrive and not taking advantage of the fact that the CNN could start processing smaller patches of the image as soon as the necessary lines have arrived, does not represent the optimal implementation either. Therefore, to avoid the first scenario and to take advantage of the second one, a different approach is proposed.

Image mode

Each video frame will be independently processed as an RGB image by the CNN implemented as the target detection stage. As the encoding method in which the video will be received is BIL, the CNN can start processing the first line of smaller patches without having to wait until the full frame is received, following in this way a pipeline strategy that favors a high throughput in the hardware implementation. Each image will be processed applying a sliding window that will move from left to right and from top to bottom with a certain stride and overlap according to the network input layer size and parameters [5], as shown in Fig. 2. This overlap will avoid missing targets when they are not totally contained within the current window. It is also remarkable that this overlap needs to be carefully chosen.

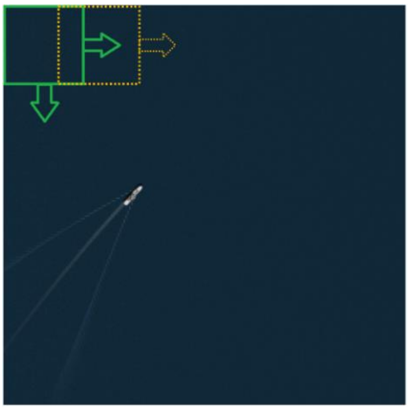

Fig. 2. Frame sweeping overview

In order to select the right window size and stride so that a complete ship can be located within the image, some calculations are needed. The sensor has an FoV of 2.5 degrees and the image used in the demonstrator will have a size of 1920x1080 pixels (i.e., Full HD resolution). Besides, the satellite will be orbiting at around 500 km of altitude. With all the above, the pixel size in the vertical and the horizontal dimension of the image acquired by the sensor embarked on the VIDEO instrument would be equivalent to 3.66 m and 2.72 m, respectively. If the Seawise Giant is taken as the longest ship ever built, with a length of 458.45 m, it would need around 169 pixels according to the previous calculations. Accordingly, it is concluded that a window size larger than ~200 pixels should be enough. Therefore, a window size of 256x256 should give enough margin for target detection.

The network result will indicate if there is at least one target present within the current window area or not. The window will be small enough to locate the targets within the image but also it has to be large enough to be able to contain most of a target so it can be correctly detected by the CNN. Finally, as soon as the result of the current window is positive, the process will stop triggering a flag that informs to the system that a target has been detected. In that case, the detection mode will stop handing over to the video mode. This process is reflected in Fig. 3. It is remarkable that video compression does not take place during *Image Mode*, since in this first scenario data are not sent to ground.

Fig. 4 presents how the validation set-up is expected to work in *Image Mode*. First, the CNN is configured through AXI4-Lite. Then, the CNN is fed with 256x256 frame blocks through a DMA connected to a FIFO to adapt data rates. Each frame block has an associated ID that will allow to know if the block under analysis returns a positive value (i.e., a target has been detected) or not. Since the image size, the window size and the overlap are known values, the number of blocks and their coordinates within the image can be calculated and stored beforehand, associating those values to each block ID. Once the target has been detected, the CNN will trigger the decision signal and it will send the block ID back to the soft microprocessor, which based on these two values will make the decision of switching to *Video Mode* or not. As it can be observed in Fig. 4, the video compression stage is not enabled in this point.

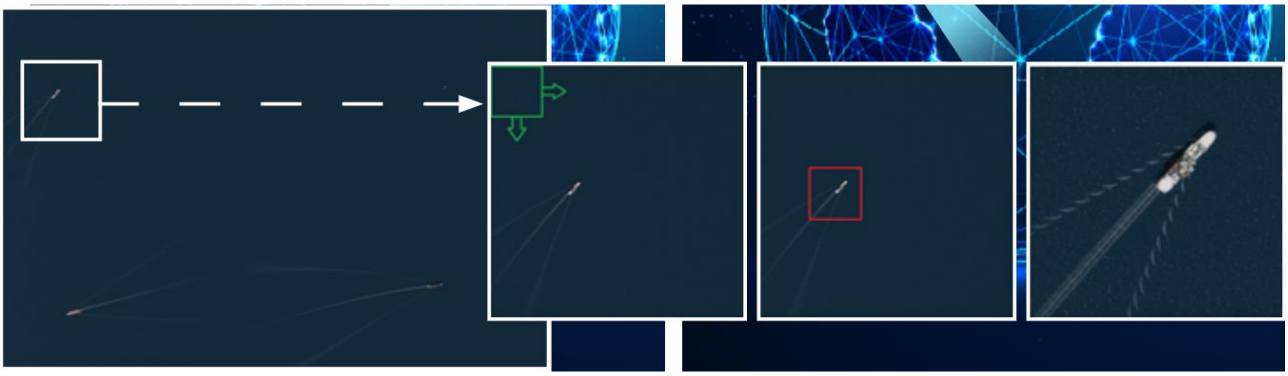

Fig. 3. Switching from Image to Video Mode

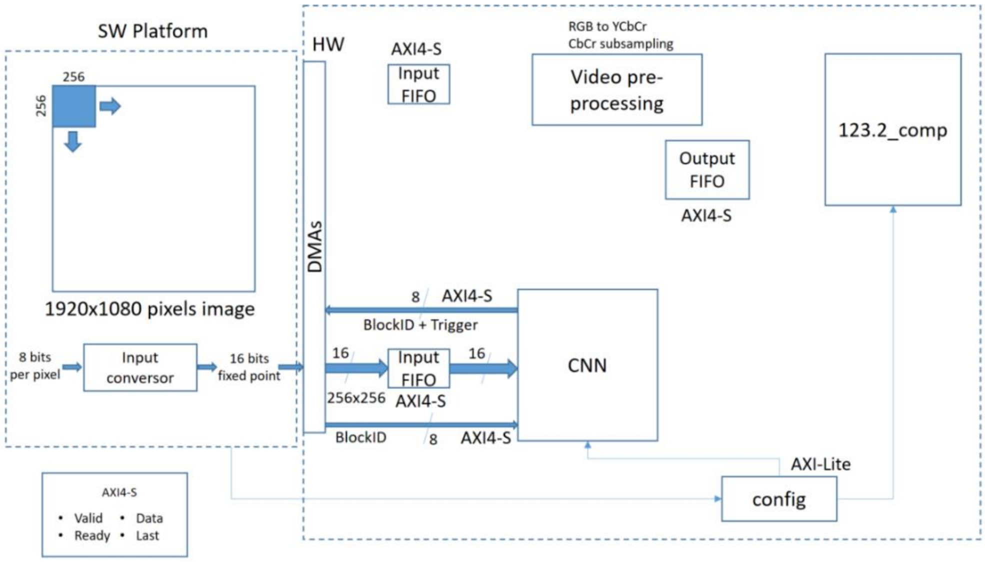

Fig. 4. System overview in Image Mode

If *Video Mode* is triggered, the video compressor is properly configured, while the CNN now performs a tracking of the target. Based on the ID of the 256x256 frame block that the CNN has detected as ‘positive’, the microprocessor will conform a 512x512 sub-image using the surrounding 128 pixels of that block. That new sub-image will be sent to the input FIFO of the CCSDS 123.0-B-2 compressor through an extra DMA.

Then, 256x256 blocks extracted from the same sub-image plus each block ID are also sent to the input FIFO of the CNN in this mode, moving again the 256x256 window from left to right and from top to bottom with an overlap of 128 pixels. Once the CNN has detected where the ship is within the sub-image, it will send the block ID back to the microprocessor so that this information can be forwarded to the CCSDS 123.0-B-2 compressor. With the entire sub-image plus the block ID, the compressor will be able to process the image applying lossless (or near-lossless with a low error limit) compression to the 256x256 region pointed by the CNN, thanks to the block ID and the stored coordinates, and lossy compression with a high error value around that area. It is remarkable that this processing strategy implies that in *Video Mode* the compressor will work one sub-image behind the CNN, as the CNN will have to process that sub-image first to let the compressor know where to apply lossless compression (i.e., where the target is in that sub-image). Fig. 5 shows how the *Video Mode* is expected to work.

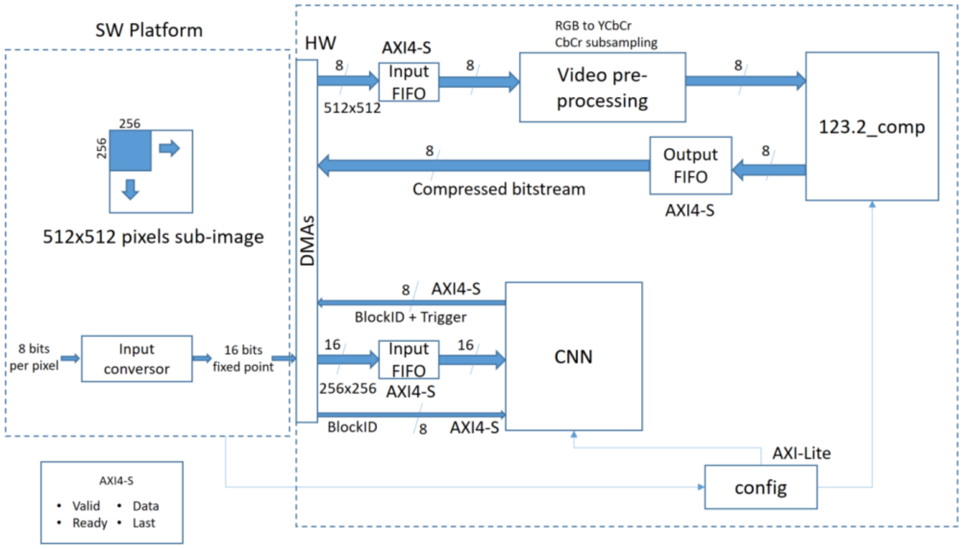

Fig. 5. System overview in Video Mode

Preliminary results

The whole video processing chain is currently under validation. A test set-up has been developed on a Xilinx Kintex UltraScale XCKU040 FPGA, allowing to transfer it then to the space-grade XQRKU060 FPGA. This set-up includes, in addition to the video processing engine itself, a MicroBlaze embedded microprocessor for system initialization and configuration; the AXI infrastructure to interconnect the different modules among them; and multiple DMAs to handle data transfers between the system and the external memory, where the video sequences are loaded and accessed through a dedicated DDR4 memory controller. The whole validation set-up also runs at a clock frequency of 100 MHz, except the DDR4 controller, which is fed with a dedicated clock at 300 MHz for high-speed accesses to the external RAM.

A test dataset has been created in the context of the VIDEO project, comprised by a collection of short video sequences where boats are captured at different locations (e.g., offshore, near the coast, at the port) [8]. This dataset is used for both CNN characterization and video compression stages.

The test procedure is managed from a control and monitoring PC, which starts data transfers with the external RAM located in the Xilinx KCU105 Development Kit. Each time the video mode is triggered and a compression is finished, three outputs are generated (one per compression instance) and sent back to the control PC, which are then independently decompressed by using an in-house software developed in C++ and compliant with the CCSDS 123.0-B-2 standard. The next step is to merge the three decompressed files into a single YCbCr video sequence, finally performing an inverse colour conversion from YCbCr to the RGB color space for visualization purposes.

The resources utilization of the whole processing chain is summarized in Table 1. These results are for an implementation able to handle RGB video segments up to 512x512 pixels. Although higher video segments could be processed, this size is selected because in the VIDEO processing chain the compressor is preceded by a target detection stage that works with this frame size, as explained in [5]. In addition, higher is the video segment to be compressed, higher is the memory utilization of the video processing chain. The luma compressor uses the 3 previous frames to perform the inter-prediction, while the chroma compressors only employ the previous one. However, the number of previous frames used to perform the inter-prediction stage for the luma channel can be reduced, if needed, to decrease the internal memory utilization, without compromising the performance results in terms of rate-distortion ratio.

The total logic resources utilization is around the 35.5% and the 75% of the available registers and LUTs, respectively), while almost the 82.4% of DSPs are used, mainly by the calculations performed by the CNN in 16-bit fixed point. The video converter employs almost the 4% of the consumed DSPs to perform the multiplications to convert from RGB to YCbCr color space, while the compression instances are around the 2% each one.

Regarding frame rate results (measured in FPS), the whole processing chain works always over 10 FPS for a maximum video resolution of 1920x1080 pixels (i.e., HD video quality). These results accomplish the performance objectives and the defined schedule of the VIDEO project, developing a functional prototype in a relatively short time (around 8 months). However, obtained results could be even better, since the main data dependencies present in the processing chain are not properly alleviate by the HLS tool, though some tasks can be performed in parallel. This prevents from achieving the maximum theoretical performance.

Table 1. Resources utilization of the VIDEO processing chain on Xilinx Kintex UltraScale XCKU040

Module	BRAMs	DSPs	Registers	LUTs
Detection network	229.5 (38.32%)	1405 (73.04%)	134284 (27.55%)	136960 (56.53%)
Video compressor	322 (53.6%)	176 (9.3%)	37557 (7.9%)	44660 (18.4%)
TOTAL	**551.5 (91.92%)**	**1581 (82.34%)**	**171841 (35.45%)**	**181620 (74.93%)**

CONCLUSIONS

This work presents the H2020-funded Video Imaging Demonstrator for Earth Observation (VIDEO) project, which has the main objective of developing a next-generation instrument for Earth Observation. In this project, in which a total of 6 European entities participates, leaded by Thales Alenia Space in France, the ULPGC/IUMA is responsible of the design and development of the whole video processing chain, which includes object detection (mainly ships, the use case of the VIDEO project), tracking and video compression of a wide scene with a high resolution.

To accomplish the VIDEO project requirements, a CNN was developed for detection and tracking purposes, while video compression is performed by using an extended version of the CCSDS 123.0-B-2 compression algorithm, originally though for hyperspectral images, to be able to manage video sequences. Currently, the video processing chain is under validation, joining both stages in the same FPGA to work together. Preliminary results demonstrate the feasibility of this solution to be implemented on hardware available on-board satellites with the required performance. Next steps are related to the finalization of that functional prototype to then improve the performance in terms of FPS.

ACKNOWLEDGEMENT

This work was supported by the European Union's Horizon 2020 Research and Innovation Program through the Video Imaging Demonstrator for Earth Observation (VIDEO) Project under Grant 870485.

REFERENCES

[1] A. D. George and C. M. Wilson, "Onboard processing with hybrid and reconfigurable computing on small satellites," *Proceedings of the IEEE*, vol. 106, no. 3, pp. 458–470, 2018.

[2] R. Castano, D. Mazzoni, N. Tang, R. Greeley, T. Doggett, B. Cichy, S. Chien, and A. Davies, "Onboard classifiers for science event detection on a remote sensing spacecraft," in in *Proc. Of Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.

[3] Y. Ji-yang, H. Dan, W. Lu-yuan, G. Jian, and W. Yan-hua, "A real-time on-board ship targets detection method for optical remote sensing satellite," in *IEEE 13th International Conference on Signal Processing (ICSP)*, 2016, pp. 204–208.

[4] VIDEO Consortium, "Video Imaging Demonstrator for Earth Observation," https://video-h2020.eu, accessed: 2021-12-09.

[5] R. Neris, A. Rodríguez, R. Guerra, S. López, and R. Sarmiento, "FPGA-Based Implementation of a CNN Architecture for the On-Board Processing of Very High-Resolution Remote Sensing Images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 3740–3750, 2022.

[6] Y. Barrios, R. Guerra, S. López, and R. Sarmiento, "Adaptation of the CCSDS 123.0-B-2 Standard for RGB Video Compression," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 1656–1669, 2022, doi: 10.1109/JSTARS.2022.3145751.

[7] Y. Barrios, R. Guerra, S. López and R. Sarmiento, "Performance Assessment of the CCSDS-123 Standard for Panchromatic Video Compression on Space Missions," in *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1-5, 2022, Art no. 5507905, doi: 10.1109/LGRS.2021.3099032.

[8] A. P. Antonio-Javier Gallego and P. Gil, "Automatic ship classification from optical aerial images with convolutional neural networks," Remote Sens., vol. 10, no. 4, 2018, Art. no. 511.