International Doctorate School in Information and Communication Technologies

DISI - University of Trento

ULTRA-LOW-POWER VISION SYSTEMS FOR WIRELESS APPLICATIONS

Nicola Cottini

Advisor:

Ing. Massimo Gottardi

Fondazione Bruno Kessler

Abstract

Custom CMOS vision sensors could offer large opportunities for ultra-low power applications, introducing novel visual computation paradigms, aimed at closing the large gap between vision technology and energy-autonomous sensory systems. Energy-aware vision could offer new opportunities to all those applications, such as security, safety, environmental monitoring and many others, where communication infrastructures and power supply are not available or too expensive to be provided,

This thesis aims at demonstrating this concept, exploiting the potential of an energy-aware vision sensor, developed at FBK, that extracts the spatial contrast and delivers compressed data. As a case study, a custom stereo-vision algorithm has been developed, taking advantage of the sensor characteristics, targeted to a lower complexity and reduced memory with respect to a standard stereo-vision processing. Under specific conditions, the proposed approach has proven to be very promising, although much work has still to be done both at sensor and at processing levels.The last part of this thesis is focused on the improvement of the custom sensor. A novel vision sensor architecture has been developed, which is based on a proprietary algorithm, developed by a partner of FBK and targeted to surveillance applications. The algorithm is based on adaptive temporal contrast extraction and is very suitable to be implemented at chip level. Although the output of the algorithm has strong similarities with the spatial contrast vision sensor, it relies on temporal contrast rather than spatial one, which is much more robust for event detection applications. A first prototype of ultra-low power adaptive temporal contrast vision sensor has been developed and tested.

Keywords

Custom vision sensor, Ultra low power, Image processing, Custom algorithm, Stereo vision.

Contents

List of Figures

Chapter 1

1. Introduction

Sensors are becoming increasingly pervasive in our everyday life. With minimum dimensions and less infrastructures, sensory networks need to embed computing resources and wireless data communication maximizing their operating lifetime and minimizing their environmental footprint. Environmental monitoring is the main application area where wireless sensor networks (WSN) may have a huge impact. Air pollution monitoring, agriculture, control of greenhouses, forest fire detection and structural monitoring are only few application examples of where a large amount of data needs to be gathered by the sensors and pre-processed to be finally transmitted. In WSNs sensors shouldn't need infrastructure and should only require low maintenance. They communicate wirelessly with the network and are powered by batteries which are recharged by natural sources. This will make the sensing nodes to be energy autonomous for long-lasting operation. Currently, most of these nodes make use of single sensors such as temperature, pressure and humidity, working only intermittently and occupying a very low data bandwidth for communication. The use of more complex sensing technologies is currently closed to the WSN, due to large amount of collected data and the corresponding computing resources to be committed and related power consumption. In particular, vision is the sensing technology with the largest information density, which requires to be processed in real-time through high-performance computing platforms. The natural way obtaining information of the world around us is visual. We obtain more than 90% of information about the world surrounding us with our eyes, and about half of our brain is busy with its interpretation. Even small animals, birds and insects can easily interpret the visual world surrounding them - this with a fraction of the computational power of an ordinary computer. On the other hand, an artificial vision system is a system that observes the visual world around it and interprets it to provide information about the scene. Such systems are currently bulky, expensive, power hungry, and instead of having cognitive capabilities are often limited to image recording. They are widely used in surveillance and security systems, traffic and pedestrian monitoring, etc., which require infrastructures for the power supply and data communication. They all are based on standard electronic hardware, which is not specifically optimized for energy-aware operations. For example, a commercial imager continuously delivers sequences of images with large redundancy becouse only a small amount of the available information is used to perform a visual task. The processor is required to execute visual processing even though no relevant information occurs in the scene, turning into a large waste of power and of computing resources. These aspects are of main importance in case of a long lasting autonomous system, which has to operate with a limited available energy budget. Although microelectronic technology has brought significant improvements in system performance and energy efficiency, vision computation did not make over the years significant progress in energy-autonomous applications in recent years. While the power consumption of a standard vision system can range from a few Watts to tens of Watts, a wireless sensor node burns typically mW on average. This means that there is a gap of 2 to 3 orders of magnitude in power consumption between these two technologies.

A standard sensor platform is typically organized in a cascade of functional blocks: Sensing — A/D Conversion — Digital Signal Processing — Transceiver.

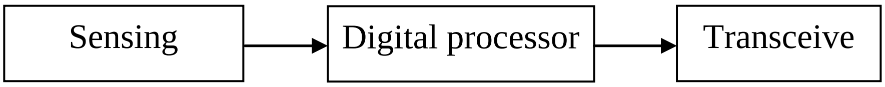


Figure 1. Simplified block diagram of a standard sensor processing system, including: sening, digital processing and communication.

As shown in (Figure 1.), the signal provided by the sensor is firstly converted into a digital form by the A/D converter, feeding the digital processor for . The output of the signal processing unit is sent to a PC or a base-station through a wired or wireless link. Here, each functional block is intended to execute a specific operation with the required performance, but has only limited interaction with the neighboring blocks. Therefore, the energy efficiency does not only depend on the performance of each single block, but it also relates to the cooperation among the different units of the system. System-level design has indeed a large impact on the energy consumption. At a first glance, embedding some programmable intelligence at sensor level, making it able to recognize and extract significant features in the scene, would drastically increase the energy efficiency of the system, without losing performance. Although, the main benefit of doing this is a reduction of redundant data, thus less data to be processed, an adaptive system would be more desirable, making it to modify its operating functions according with the specific scenarios and with the available energy. In other words, the energy management concept needs to be applied at system-level in addition to each single block of the signal path.

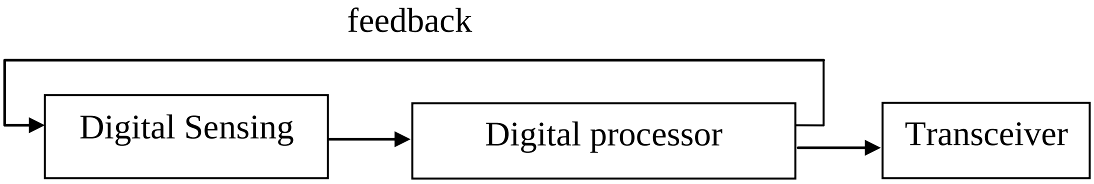


Figure 2. Simplified block diagram of a custom sensor processing system, including: digital sening, digital processing with active feedback for sensor and communication.

Algorithms for visual processing will also play a key role in the system design. As always, algorithms are severely limited by the hardware and have to be tailored to exploit the potential of the current system architecture.

CMOS technology allows the integration of image sensing with massive parallel visual processing architectures, where the custom vision sensor can be dynamically programmed though a feedback of second level algorithm (Figure 2.). This approach offer a unique opportunity of introducing novel energy-aware computational paradigms, closing the gap between vision technology and energy-autonomous systems.

This thesis aims to deal with the energy-aware visual computation issue by exploiting the potentials of two custom low-power vision sensors, combined with lightweight algorithms, based on event-detection and are targeted to monitoring applications.

Novel processing paradigms has been investigated, aimed at optimizing the senor custom data coding with minimum use of memory and computing resources. Two, proof-of-concept demonstrators has been developed, demonstrating the validity of the proposed approach.

The presented work wants to be the base for further research investigation aimed at closing the gap between vision sensor technology and energy-autonomous sensory systems.

Chapter 2

2. State of the Art

Since last decade, a lot of effort has been spent by the research community in bringing vision systems into the battery-powered category, by using Components Off The Shelf (COTS). This has brought to the development of some vision prototypes performing relatively simple event detection operations with hundreds of Watts to few Watts of power consumption. This hampers those systems to be used for long-lasting operation powered with batteries.

In fact, standard components are general purposes devices targeted to a wide range of applications. In particular, commercial image sensors are targeted to multimedia applications where, image quality and resolution are the main figures of merit. For these applications, an imager burning 50-80 mW is claimed to be a low-power component. If we consider that a wireless sensor node burns typically mW on average, we can understand that a radically different approach needs to be adopted based on custom components optimized for low power performance. In order to better identify the low power vision issues, in the first part of this Section, we will present an overview of the most popular low power vision prototypes based on COTS and custom sensors. In the last part of the Section, we will focus on some examples of ultra low power custom vision sensors.

2.1. System

Most of the groups that develop CMOS imagers today perform image processing on frame-based hardware [1][2][3][4][5][6][7][8]; this means that the image is acquired during exposition time and dispatched to the output of the device, pixel after pixel, sequentially, in a raster-scan fashion. In most cases, this is the only feasible approach, driven by the market availability of the imager technology.

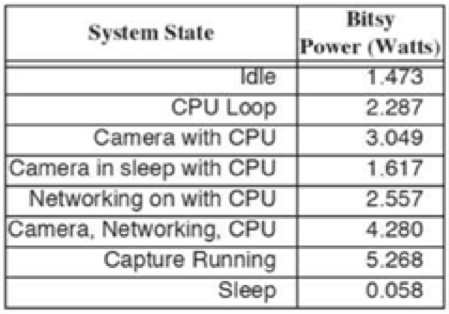
Commercial imagers (Components Off the Shelf: COTS) operate on frames. Acquisition of frames and processing them dominates the power consumption of existing demonstrations.

In the field of wireless sensors networks, several video sensor nodes have been reported, which are connection points of a network capable of sending, receiving, or forwarding information over the communications channel. All these system are based on COTS to meet the tight cost constraints typical of distributed sensing applications. For example, Panoptes [9] developed in 2003 by Feng et al. at the OGI school o Science of Portland State University and subsequently improved at the Department of Computer Science of Portland State University.

The prototype is equipped with an Intel StrongARM processor, a Logitech 3000 camera and Linux OS. Its power consumption is more than 5W Figure 3. Delivering the stream of image at 20 fps with a resolution of 320 x 240 Pixels. The new prototype adopt the Crossbow Stargate Platform decreasing power consumption at 4W.

To supply solar power to this device, a solar panel of about 2 sq. meters and a car-sized supporting battery would be required.

The elaboration executed inside consist in a decompression image coming from the standard camera feature extraction through the redundancy elimination. This permit to reduce the power consumption and the transition bandwidth.



System State	Bitsy Power (Watts)
Idle	1.473
CPU Loop	2.287
Camera with CPU	3.049
Camera in sleep with CPU	1.617
Networking on with CPU	2.557
Camera, Networking, CPU	4.280
Capture Running	5.268
Sleep	0.058

Figure 3. Panoptes, smart camera mote architecture hardware and power consumption

Another example of a visual sensor is the MeshEye [10] (Figure 4. which has much lower power consumption. Although a benchmark has to be considered to properly evaluate the performance of the system, this smart camera is claimed to last about 22 days with 2 AA batteries (less than 2 fps). This system is intended to work in periodic poll intervals, where the microcontroller wakes up periodically, acquires an image and determines if something has entered the scene. Once an object has been detected, intermediate level processing extracts its descriptive representation. MeshEye is capable of detecting, acquiring and tracking objects entering the scene, thus it is suitable for surveillance applications.

The authors claim that, a frame rate of at last 10 fps, in this configuration the life time of batteries in less than two days.

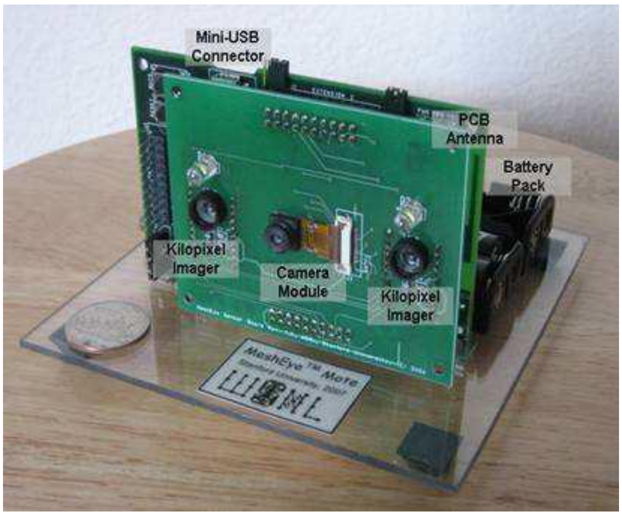
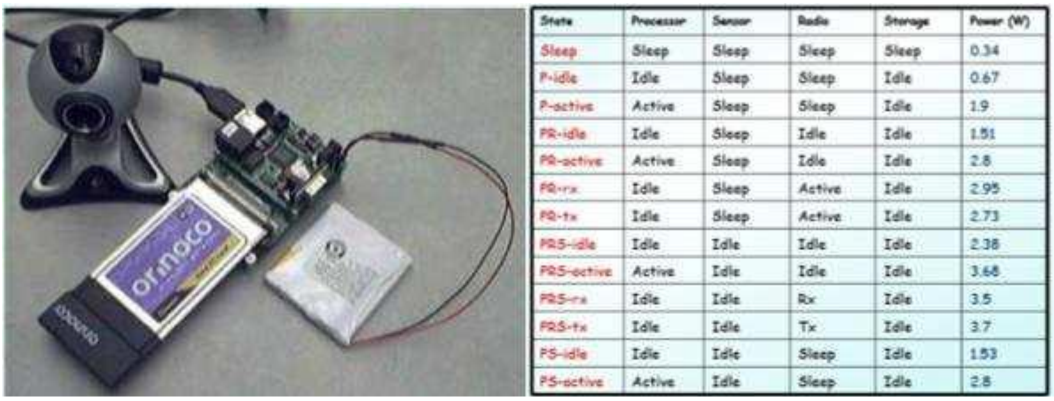


Figure 4. Mesheye, energy-efficient smart camera mote architecture.

This system incorporates 2 1k-pixel imagers (optical mice sensors) and 1 VGA resolution image sensor with a microcontroller and Zigbee wireless radio interface.
The philosophy of this system consist to use the low level imager to detect the basic movement preset in the scene and when a movement are detected the second level imager is wakeup.



State	Processor	Sensor	Radio	Storage	Power (W)
Sleep	Sleep	Sleep	Sleep	Sleep	0.34
P-idle	Idle	Sleep	Sleep	Idle	0.67
P-active	Active	Sleep	Sleep	Idle	1.9
PR-idle	Idle	Sleep	Idle	Idle	1.51
PR-active	Active	Sleep	Idle	Idle	2.8
PR-rx	Idle	Sleep	Active	Idle	2.95
PR-tx	Idle	Sleep	Active	Idle	2.73
PRS-idle	Idle	Idle	Idle	Idle	2.38
PRS-active	Active	Idle	Idle	Idle	3.68
PRS-rx	Idle	Idle	Rx	Idle	3.5
PRS-tx	Idle	Idle	Tx	Idle	3.7
PS-idle	Idle	Idle	Sleep	Idle	1.53
PS-active	Active	Idle	Sleep	Idle	2.8

Figure 5. Meerkats Power-aware, Self-Managing Wireless camera network for Wide Area Monitoring. and Power consumption

Meerkats [11] (Figure 5. belongs o the same class of Panoptes, but uses more recent components (XScale processor). The goal of this work is Process an image before transmission, cut off a region involved by event, and extracting features such as motion flow, may decrease the amount of data being transmitted.

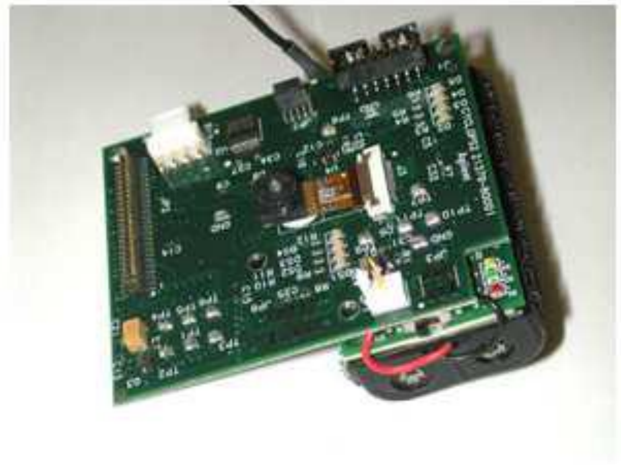


Figure 6. Cyclops couples with a Berkeley Mote and they represent a wireless vision network node.

Cyclops (Figure 6. is a much lower power device, making use of Xilinx CPLD and Atmel microcontroller ATmega128L and a sensor with a CIF resolution (352×288) [12]. It achieves less than 4 fps for basic (presence/absence) object detection task on small images (128 x 128 pixels).

Power Consumption: a) stand-by < 50µA; b) 5mA@4MHz clock freq.; c) 11mA@13MHz clock freq.

The wireless node implemented by Ferrigno [13] is equipped with the Microchip PIC16LF877 microcontroller and performs software image compression at less than 1 fps (Figure 7. . This system is intended to acquire one-shot images and to transfer them wirelessly at low data rate. It doesn’t perform local image processing, thus is not suitable for surveillance applications.

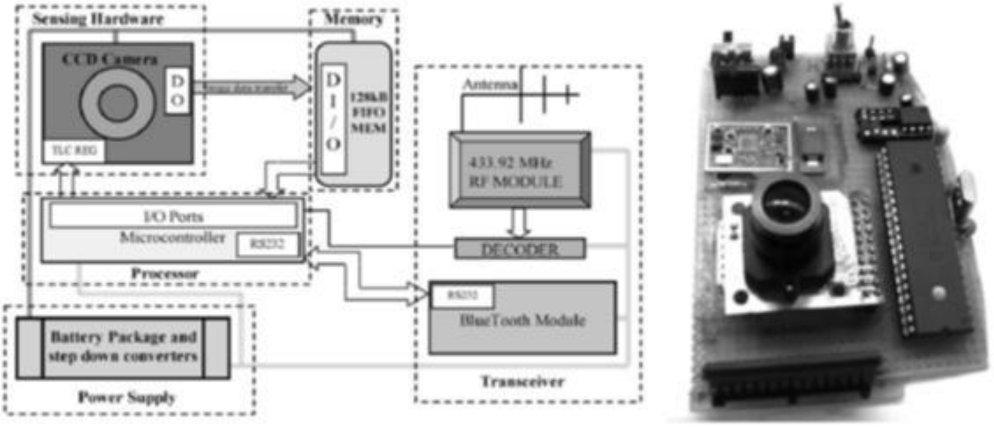


Figure 7. Visual sensor node for Bluetooth-Based Measurement Networks.

In a surveillance application, the sensor may send alert images for verification to security personnel only when the sensor identifies an alert condition. In monitoring applications (people counting, traffic monitoring) no images may be sent, only small data packets requiring a minimal bandwidth and minimal power for transmission.

In contrast with standard systems, would not be necessary to deliver continuously stream video information. Ideally, triggers on significant events in the scene, extract features, and dispatch only this information, to permit drastically cutting down the bandwidth and of course energy consumption.

In order to do this is necessary to understand where removing the most irrelevant and redundant features from the data in efficient way. The most efficient way to extract the important information present in the scene it’s execute this pre elaboration directly on chip avoiding an overload of data to other elaboration blocks. This approach require custom sensors, because in general the sensors are not able to deal with events.

One of the main example of system based on custom sensor we presented by Texeira [14]. He proposed a camera sensor for behavior recognition based on a important data coding named the AER (Address Event Representation).The AER mimics the methods of transmitting information through spikes trains, proper to the optic nerve, more generally, of the majority of neurons.

The AER is a communication protocol in frequency. The information is encoded by the intervals of time inter spike (ISI, Inter Spike Interval). The pixels used in this retina allow to best use the capacity of channel. Only the pixels that have something to communicate requesting to external bus access. Completely different situation respect to a raster used by conventional sensors. The sequence of spikes that travel along the external bus are only those from the pixels that show variations in light intensity.

The pixels of the matrix are completely independent and no busy signal is communicated to the various neurons in the case where one of them is issuing the spike, their loading the external bus with a identifies code (x, y locations). A selection system between neurons in the competition for access to the bus is necessary. The circuit part which operates this selection is called the arbiter. The arbitrator chooses which neurons give access to the bus AER depending on the timing for submission of requests access.

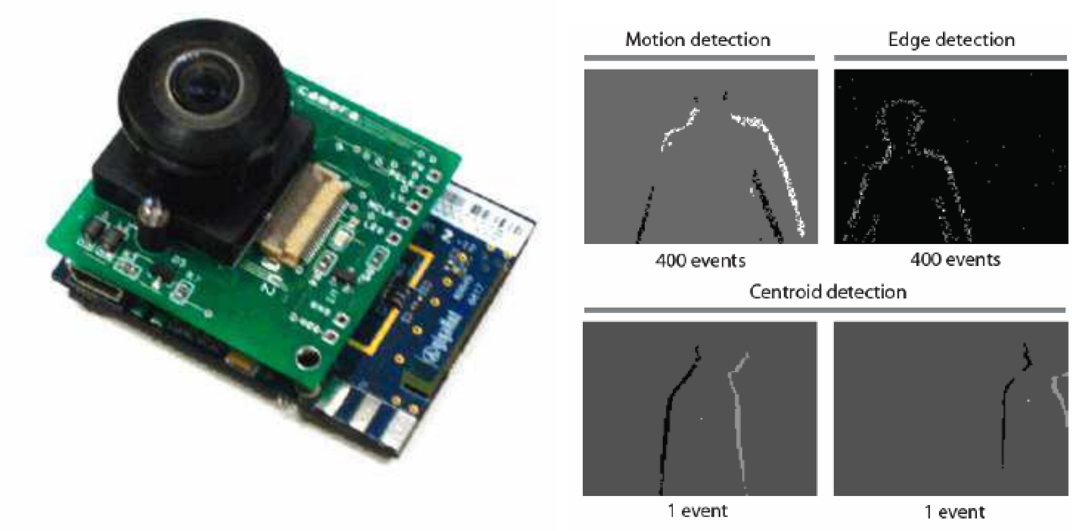


Figure 8. Bio-inspired vision node based on address-event image approach.

The system aims at demonstrating the benefits of using an Address-Event representation approach in the visual processing path. The system uses a 44mW color

VGA camera, OmniVision OV7649, interfaced with a XScale processor (PXA271). As wireless node, an Intel iMote2 has been adopted. Working in full active mode at 104MHz and 8fps, the entire system consumes 322mW in which the iMote2 is responsible of 279mW.

The Anafocus Eye-RIS [15] is an interesting example of vision systems on the market, which is based on a custom vision chip [16]. The Vision System on-Chip (VsoC) architecture (Figure 9.) performs image processing at three different levels:

1 : pixel-level processing, in which each pixel includes analogue and binary processor and memory.

2 : column-level processing which represents a linear array processor able to readout and cooperate to process one or several image rows as required by the algorithm.

3 : system-level processing is a powerful on-chip microprocessor designed to speed-up power consuming task.

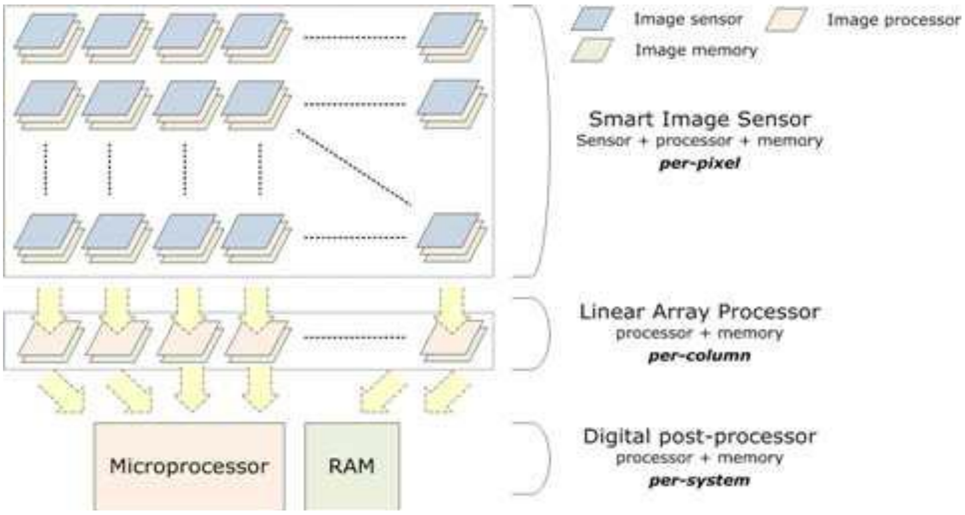


Figure 9. Architecture of the Anafocus Eye-RIS system.

In this way image processing is split in two steps: an image pre-processing and an image post processing.

Image pre-processing targets extracting useful information from the input image flow; this means eliminating all redundant, and therefore useless, information for the specific algorithm being accomplished. It consists of relatively simple processing tasks, such as image convolutions, spatial filtering, morphological and statistic operations; combined in algorithms that are intensively applied to each captured image. Image post-processing targets making complex decisions and supporting action-taking. It normally involves complex algorithms within long and involved computational flows and may require larger accuracy than early processing. The major benefit introduced by this architecture is the reduction of on-chip memory and overall system power consumption.

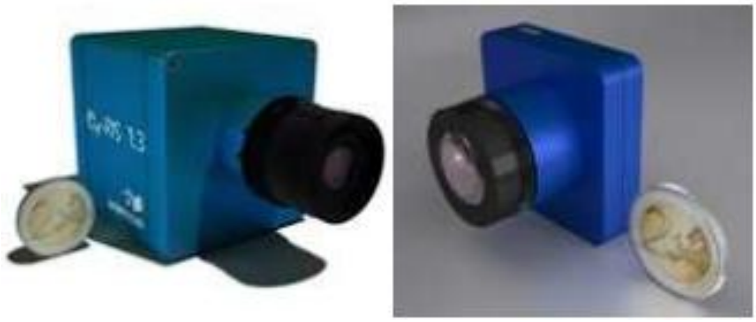


Figure 10. Eye-RIS v1.3 and v2.1 vision system.

The Eye-RIS v1.3 (Figure 10.) vision system has a resolution of 176x144 pixels and performs high performance image acquisition at high speed (over 10000 fps) with mixed-signal image processing with fast electronic global shutter characteristic, showing a typical power consumption of 1.5W.

The multiple board architecture and reconfigurable FPGA allows the system high flexibility, permitting its easy adaptation to the requirements of specific applications, but these approach introduce some limitations in term of power consumption.

The Eye-RIS v2.1 is a more compact system with a lower power supply and power consumption (700mW).

The last two works demonstrate the importance to optimize all parts of the system, but especially the vision sensor. In fact it is just the sensor to determines the type and the quantity of date will be processed. It therefore becomes necessary to analyze the state of the art of custom sensors.

2.2. Vision sensors

Current commercial imagers are almost always designed for multimedia applications - mobile phones, digital video cameras and toys, where low cost and high image resolution are the main figures of merit. CMOS has almost replaced CCD technology in the imagers scenario, at least for the consumer market. The big advantage of CMOS image sensors over the traditional CCD sensors is in their capability of integrating sensors, A/D conversion, digital signal processing, such as auto exposure control, pixel correction, face-smile detection etc., in a true System On Chip paradigm.

On the other hand, CMOS offered the possibility of developing novel architectures and un-conventional approaches for image sensors, which are more oriented toward custom and special applications requiring sensors with advanced performance.

Although the concept of “vision sensor” is a fairly mature term, which was introduced few decades ago (ref Carver Mead), the advent of the CMOS sub-micron technology has made possible the integration of increasingly complex tasks, enhancing the potentials of custom sensors in several application scenarios.

Among others, low power performance is going to become a priority. This is because mobile devices are almost ubiquitous as well as sensors is the technology which promises to drive the semiconductor market for the next decade.

With such common architectures, sensor power consumption is secondary to that of the overall system: DSP, memory and communication unit. All this results in high power consumption. For example, a 70mW commercial VGA CMOS imager is claimed to be an ultra-low power sensor. Powered with a small 950mAh Li-ion battery, the imager can only run for 38 hours, without taking into account additional system components, which are usually much more power hungry than the sensor itself. This gives us a rough estimation of the lifetime for a battery-operated vision system, based on commercial components, even in the most optimistic scenario. Several tentative battery-powered vision systems, based on Commercial-Off-The-Shelf (COTS) components, have been produced in the last decade. They did not obtain encouraging results in terms of operating lifetime, due to their large power consumption, ranging from hundred mW up to several Watts. On the contrary, they proved that the custom design is definitely the best approach for developing low-energy vision systems.

In the literature, several examples of custom vision chip implementations are reported, aimed at low-power applications. While interesting, the majority typically confront only the sensor perspective, without taking into account other important system-level issues: high-level image processing, data communication and energy management.

Specification	Kagawa[17]	Gottardi[18]	Fu[19]	Hanson[20]	K.Cho[21]	Tang[22]	Law[23]
Cmos Technology	0.35µm	0.35µm	0.5µm	0.13µm	0.13µm	0.35µm	0.35µm
Number of Pixel	128x96	128x64	64x64	128x128	128x128	128x96	32x32
Fill Factor	18.5	20	23	32	38	39	21
Frame Rate	9.6fps	Up to4000fps	60 fps	8.5 fps	15 fps	9.6 fps	21 fps
Supply Voltage	1.35V	3.3V	3V	0.7V	1.25V	1.35V	1.5V
Total Power	55µW@9.6fps	20µW@10fps	1.2mW	0.7µW@0.5fps	N/A	55µW@10fps	N/A
Consumption	460pW/fr.pix	269pW/fr.pix	4.9nW/fr.pix	85pW/fr.pix	N/A	460pW/fr.pix	821pW/fr.pix

Table 1. Lists the most recent low-power image sensors.

Interestingly, very advanced performance is claimed. However, it is worthwhile to analyze carefully the architectures reported. This is not always a simple task. In fact, the presented implementations are very different from each other (pixel topology, ADC, chip interface, subsequent processing). Moreover, the data presented on power consumption is not homogeneous: total power, power/pixel, power/array, power/ADC conversion, etc. Although no standard has been defined yet to enable rigorous comparison between the implementations, a good figure of merit is the power consumption per frame per pixel (W/frame.pixel). This value includes the actual consumption of the pixel together with a share of consumption of the sensing circuit, A/D conversion and chip interface.

State-of-the-art CMOS imagers exhibit pixel size of 1,4µm x 1,4µm and 8 Mega-pixel resolution [24].

Recent trends toward wireless sensor networks necessitate an efficient way to extract visual data from a camera meeting the limited energy budget of the sensor node. Conventional scanned imagers are not able to fulfill these requirements due to their poor efficiency in the use of the signal bandwidth and the requirement for expensive video processing on the raw pixel data.

Sensor network nodes are limited by power, computation and communication capabilities. For this reason it is important to use sensors that collect only the necessary information in a scene. There is a need to limit the use of resources during operation of the sensor network, especially energy expenditure, which is related to the node lifetime and ultimately its usefulness. Communication is costly as a result of lifetime constraint, since the radio is the most power-hungry component in the node, for this reason it is important to reduce the information sent as much as possible.

An interesting approach in visual data communication is represented by the implementation of event-driven communication systems.

This approach was taken by Teixeira et al. [25] at Yale University, New Haven, CT, USA. Their work proposes a non-standard imager in which the concept of frame is replaced by an *Address Event Representation (AER)*.

The *(AER*[26][27][28]) has been demonstrated to be an efficient method to communicate information among bio-inspired subsystems, especially for vision chips. Several implementations have been reported in the literature [29][30][31]. AER systems are based on spikes generated by those pixels reaching threshold. The pixel generating a pulse asks the system to be read out. The communicating system is asynchronous and assigns resources only on demand, resulting in better energy efficiency than traditional synchronous systems, which allocates the same bandwidth to all the pixels of the array. A reported 64 x 64 pixel image sensor adopts the AER architecture with a low power consumption of 5.75µW and a dynamic range of 235 dB [32]. Even though very wide dynamic range has been demonstrated, the imager is not intended to perform image preprocessing.

In literature, several examples of vision sensors implementations are reported, targeted to low power applications. Even though most of them represent interesting implementations, the majority approaches the problem from the sensor perspective, without taking into account the system-level data communication and energy issues [33][34][35][36].

In fact, almost all the developed vision sensors are more concerned about performance rather than performance/power consumption.

In the next paragraphs, we will analyze few examples of vision sensors embedding different image processing algorithms which are targeted to the anaysis of the activity in the scene (motion and/or scene changes).

Temporal gradient

The vision sensor present by Tobi Delbruck [37] adopt the AER approach where the output consists of asynchronous address-events that signal scene reflectance changes at the times they occur. This sensor is inspired to biological retina principle. The figure Figure 11. represent the output of the sensor it possible to show the efficient filter able to remove the background activity, this permit only the movement extraction.

The sensor is characterized by 128x128 Pixel by it has 40 x 40 µm$_2$ pixels with 9.4% fill factor, the Dynamic range is 120 dB and chip power consumption is 23 mW.

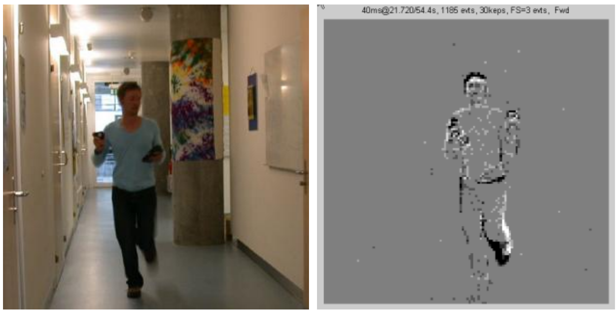


Figure 11. Dynamic scene taken by Tobi Delbruck's sensor

Etienne-Cummings [38] propose A 189 x 182 Active Pixel Sensor (APS) for temporal difference computation fabricated in 0.5 micron CMOS process, contains in-pixel storage elements for a previous image frame. The chip consumes 30mW at 50 fps from 5V power supply 8-bit precision with fill factor of 30%

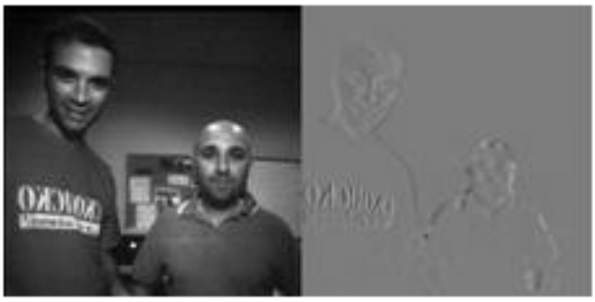


Figure 12. Sampled intensity from the Temporal difference about Etienne-Cummings sensor.

Other publication are focused on power consumption, Dongsoo Kim [39] declare 1 mW with a 3-V of power in active state where the sensor it is able to compute the temporal difference between continuous frames and filter out redundant data.
The sensors is caraterized by 64x64 pixel it has Each pixel occupies an area of 29 × 28 μm2 with a fill factor of 23%.

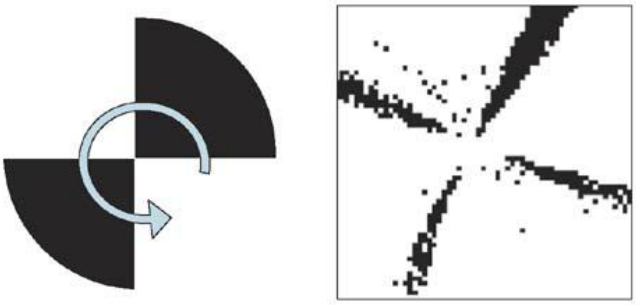


Figure 13. Dongsoo Kim Image sensor, test board and measured results with the human movements.

The sensors realized by P. Lichtsteiner and T. Delbruck [40] reduces image redundancy by responding only to temporal changes in log intensity. Where Static scenes produce no output. Image motion produces spike event output that represents the changes in image intensity. It has 64x64 pixels Each (40um^2) the fill factor (8.1%) where the power consumption is 7mW.

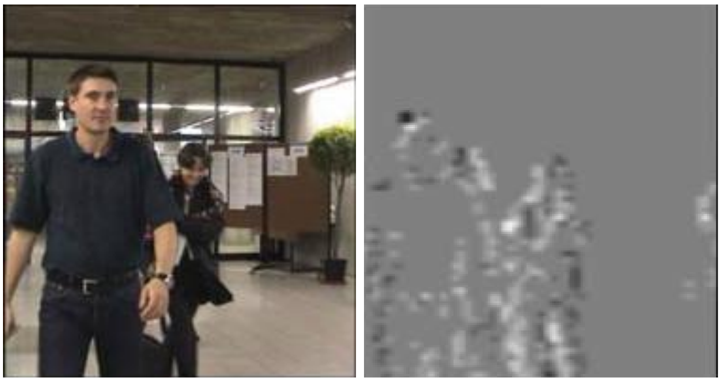


Figure 14. Lichtsteiner’s output, where the two people are moving to opposite direction . The leading edges produce OFF spikes and trailing edges produce on spikes.

Not all custom sensors are based on a temporal different but an alternative it’s the spatial gradient extraction.

The 64x64 bio-inspired pixels vision sensor developed by FBK with adaptive dynamic background subtraction the sensor detecting temporal changes in light intensity between two successive frames in binary form.

This sensor can dynamically adapt to changing scenes, in order to compensate for slow-varying levels of illumination and detect the high-varying. This sensibility can be tuned by the external control. We will see better this chip in the next chapter.

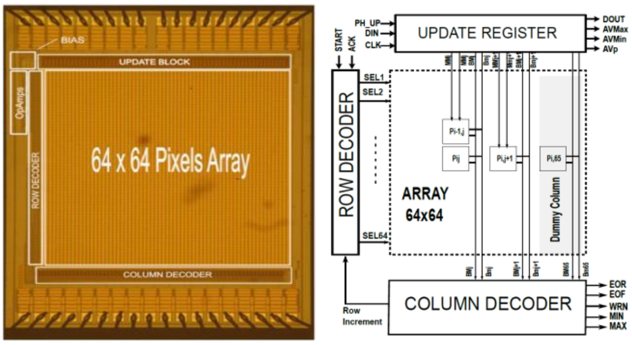


Figure 15.   Bio-inspired vision sensor developed by FBK

Other types of elaboration can be Implemented on-chip one of this are the spatial filter, are used for feature extraction such as edge detection.

Below we report some spatial gradient implemented on chip present in literature.

Spatial gradients

An important example of a visual sensor based on spatial gradient is a chip realized by Rüedi [41]128 x128 Pixel with 120-dB Dynamic-Range. The vision sensor delivering the spatial gradient magnitude and direction of image features, where the Contrast direction is a important information for performing recognition operations.

The chip dispatch information by decreasing order of contrast magnitude obtaining two main advantages first all the significant information is delivered first and the amount of data dispatched out of the circuit can be tuned for different task. But the relevant number of operation inside the pixel increase the pixel size is 69x 69 µm^2 and reduce a fill factor of 9%.

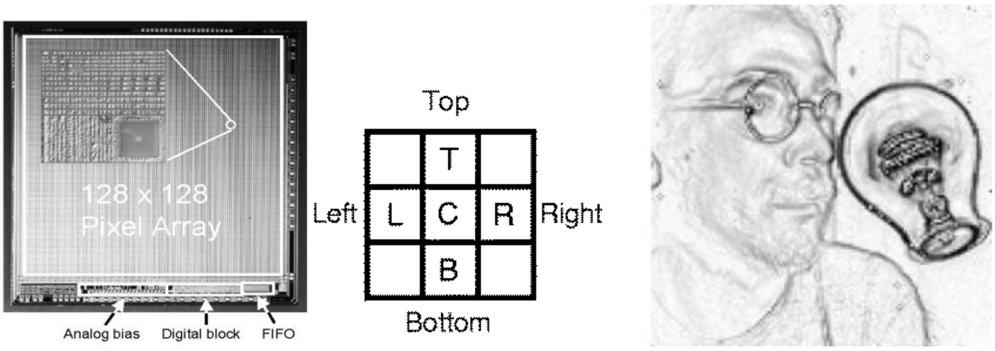


Figure 16. Rüedi's sensor, contrast kernel, masch, contrast representation

Another example is the sensors developed by Dongsoo Kim [42] characterized by pixel area is 16×21 µm^2 and the power consumption performance is not very low power, is 1.2 mW at 3 V .

The 128 ×128 smart pixel array extract intensity, spatial contrast, and temporal difference images. The spatial contrast, where the pixel (i,j) finds the maximum and the minimum photo-integrated signals with the winner-takes-all (WTA) and loser-takes-all (LTA) in the 4 adjacent pixels {(i,j),(i+1,j),(i,j+1),(i+1,j+1)}.

The pixel transfers the maximum and minimum signals extracted to the column readout circuit. The column readout circuit evaluates the difference between the maximum and the minimum signal and generates an event by comparing the difference with a threshold, it means that a contour (edge) was found.

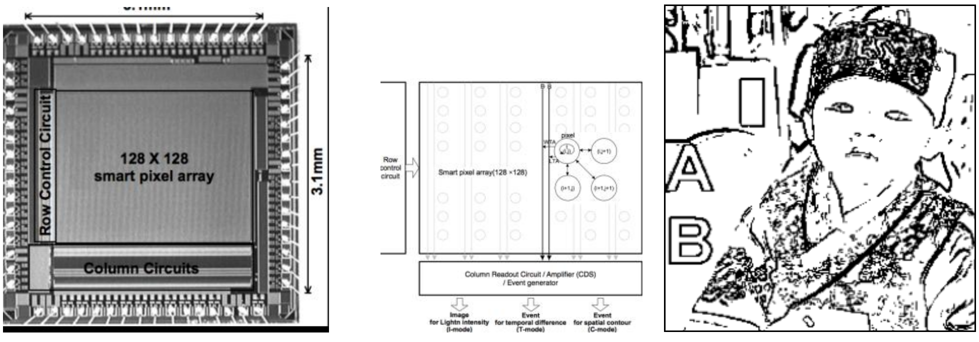
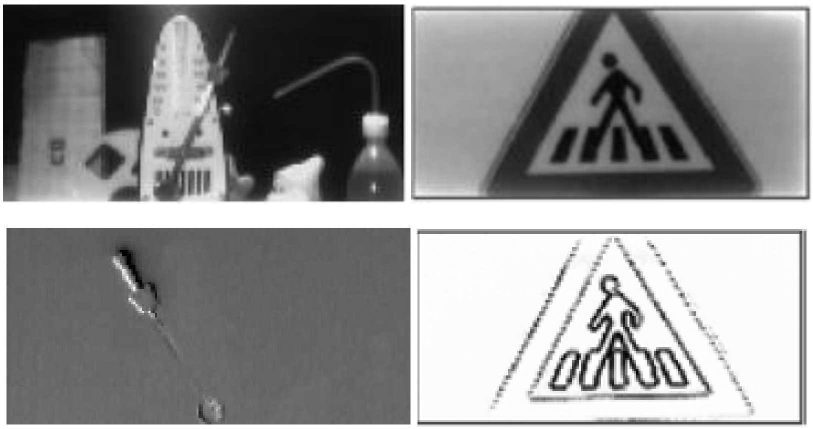

3.1mm
Row Control Circuit
128 X 128
smart pixel array
Column Circuits
Row control circuit
Smart pixel array(128 ×128)
pixel
WTA
LTA
B
B
(i,j+1)
(i+1,j)
(i+1,j+1)
Column Readout Circuit / Amplifier (CDS) / Event generator
Image for Lightn intensity (I-mode)
Event for temporal difference (T-mode)
Event for spatial contour (C-mode)
A
B

Figure 17. Kim’s sensor and Edge detection algorithm using WTA and LTA functions.

Spatio-temporal gradient

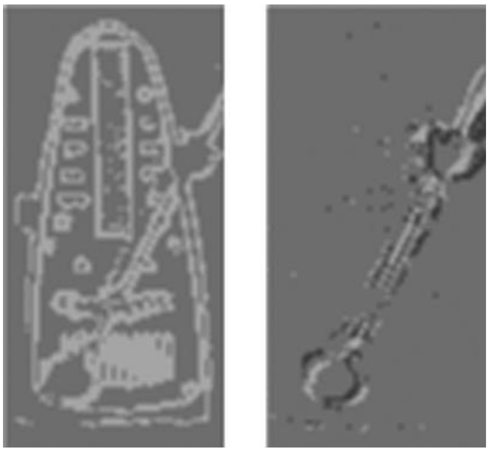
The vision sensor designed by Massari [43] consists of a 128x64 pixel array. The pixel-parallel vision sensor architecture that offers higher flexibility where novelty of the approach consists in its capacity to acquire images with a dynamic range up to about 100 dB, combining pulse-based and time-based signal processing technique, during the integration phase. The main characteristic is a highly programmable vision architecture, able to implement a different class of pixel-level spatio-temporal filtering. Spatial contrast are executed with full 3x3 pixel kernel connectivity and temporal contrast for motion detection is implemented by two successive frames difference.



(a) (b)

Figure 18. Example two different imges extraction techniques (a) Motion detection by frame difference and (b) full kernel edge detection

Gottardi [18] propose a 100 µW 128 x 64 Pixels Contrast-Based Asynchronous Binary Vision Sensor for Sensor Networks Applications. It’s a ultra-low power 128 x 64 pixels vision sensor, characterized by pixel-level spatial contrast end temporal contrast extraction and binarization.



(a) (b)

Figure 19. Example of a moving object acquired by the sensor working in (a) normal contrast mode; and (b) in motion extraction mode.

	Resolution	Pixel µm2	Fill factor	Dynamic range	Power
Tobi Delbruck	128 × 128	40 × 40	9.4%	120 dB	23mW
Etienne-Cummings	189 × 182	25 × 25	30%		30mW at 50 fps from 5V
Dongsoo Kim	64 × 64	29 × 28	23%		1 mW a 3-V
P. Lichtsteiner	64 × 64	40 × 40	8.1%		7mW
FBK temporal contrast	64 × 64	30 × 30	12%		620pW/frame*pixel
Rüedi	128 × 128	69 × 69	9%	120dB	300 mW 3.3V
Dongsoo Kim	128 × 128	16 × 21	42%		1.2 mW 3 V
Massari	128 × 64	32.6 × 32.6	24%		14 mW 3.3V
Gottardi	128 × 64	26 × 26.5	20%	100dB	100uW 3.3V

Table 2. Characteristics of the custom vision sensors presented.

Table 2 shows the main characteristics of the custom vision sensors described in this Section. In contrast to commercial components, they have relatively large pixel size due to the use of electronics for embedded processing. Although their power consumption seems to be high, compared with their poor pixel resolution, it has to be pointed out that the custom image pre-processing carried out by these sensors will in general reduce the computing resources required by the system to accomplish the specific task. This will turn into a reduction of the overall power consumption.

Chapter 3

3. Low power vision sensors at FBK

As mentioned in the previous chapter, the custom sensors are powerful candidates for energy-aware applications. Both embed custom visual processing algorithms at pixel-level with low power consumption/pixel together with pixel-level A/D conversion. Custom data coding and compression has been adopted to avoid redundancy and minimize the activity at the sensor interface.

One of the objectives of this PhD is focused to the exploitation of these two vision sensors developed at the Fondazione Bruno Kessler (Trento) and targeted to low power applications. This activity relates to the development of the electronic systems, the conception and design of novel image processing algorithms, taking advantages from the custom image processing.

FBK institute designed different type of sensor based on two different approaches:

- *Spatial contrast extraction*
- *Temporal contrast extraction*

The ***spatial contrast*** is the ability of the visual system to appreciate the contrast photometric, in other word the difference brightness of two adjacent areas. This is intended for definition as the ratio between the brightness difference of the two areas and their sum also defined as the Michelson:

$$\Delta\mathrm{I}=\frac{\left|I_{\max}-I_{\min}\right|}{\left|I_{\max}+I_{\min}\right|}$$

Often, the information is dispatched following the intensity map criterion, which is not a good representation for detecting salient features from a scene. Even though the sensor has very low-power consumption, it is not able to directly trigger salient events of the scene. Image processing is demanded outside the chip.

Another sensor develop on FBK institute is based on the temporal contrast.

The ***Temporal contrast*** is the ability to detect variations in luminance over time, is required for motion extraction.

If consider a intensity value I_{ij} to a pixel, it can be represented as:

$$\Delta\mathrm{I}_{ij}\ (t)=\mathrm{I}_{ij}\ (t)-\mathrm{I}_{ij}\ (t\text{-}1)$$

The $\Delta\mathrm{I}_{ij}$ represent the changes intensity value pixel by pixel from previous frame related to movement present in the scene, in order to discriminate real movement from noise most techniques work with some threshold.

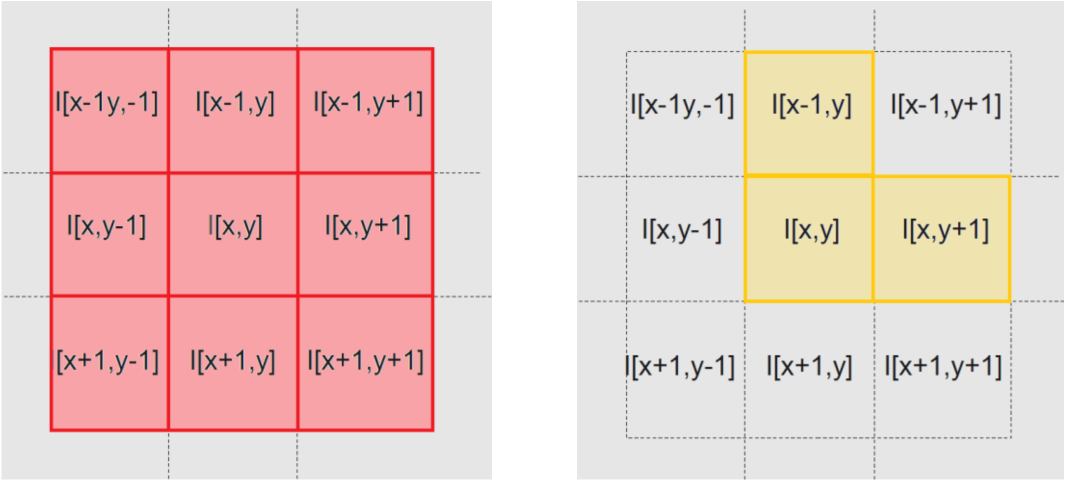
3.1. Spatial contrast sensor

One of the sensors designed by the FBK researchers Gottardi[18].This sensor directly extracts the spatial contrast of an acquired image directly on chip exploiting the Weber contrast approximation:

$$C=\frac{\left|I-I_B\right|}{\left|I_B\right|}$$

where I and I_B representing the object and the background luminance, respectively.

The spatial contrast extraction in most robust respect the standard edge extraction because the different extracted it's normalized. In this way the dynamic range of the signal analyzed is unconnected of this type of measurement, in fact it's possible to map a type of disparity in priory range known. In general this algorithm are implemented using the kernel, where the comparison between the pixels are executed.



(a) (b)

Figure 20. (a) 3x3 Kernel of standard filter, (b) Kernel of three adjacent pixel of sensor used.

The dimension of this kernel it's related of the complexity and the precision we will be obtained. The sensors for each pixel use only the different incident irradiance between a three-pixels kernel composed by, the pixel itself, the pixel on right and the pixel at its above.

The sensor at pixel-level directly extracts the spatial contrast of images in binary form through auto-adaptive technique, allowing a target to be distinguished from the surrounding background. (Figure 21.

The possibility to implement a feature extraction is obtained with Single Instruction Multiple Data (SIMD) technique exploitation. This process is complete autonomous tanks at the independent pixel implementation.

The date delivered represent only the counter edge present in the scene, in general, in the image the pixels involved are few compared with the image resolution (15%), which guarantees the minimum I/O bandwidth.

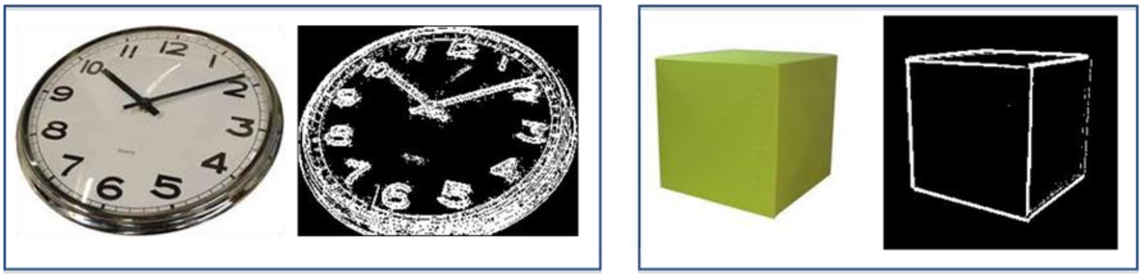


Figure 21. Simulation of the spatial contrast algorithm implemented in FBK sensor.

This hardware implementation allows to execute the elaboration directly on chip for any pixels simultaneously increasing the efficiency in term of power consumption and speed readout. Another important innovation is the address representation of the sensor, which delivers data in a sparse-matrix through a positional coding.

In the chip the image pre-processing operations are implemented through an integrated binary frame buffer, which allows the extraction of features such as contrast, extraction of the motion and the background subtraction.

The sensor has been designed to operate in two different modes:

ACTIVE: the sensor acquires and executes a temporal matching between current and reference images, dispatching the relative address pixels to the output;

IDLE: the sensor executes the same functionality, without dispatching pixels to the output, but it provide only the number of disparity pixels present in the scene.

Moreover, combining these two functionality the sensor usually stays in idle-mode, watching at the scene and estimating motion inside it, without delivering data to the output. In case the amount of change in the scene reaches a user-defined threshold, the sensor wakes-up (Active Mode) and starts delivering only the position of those pixels directly involved in the motion.

Pixels functionality: the architecture of a single pixel, is show in Figure 22.

It's composed of five basic elements: the photodiode, two comparators, a block of contrast and a memory cell to one bit, together with all other, composes a matrix capable of storing a complete image.

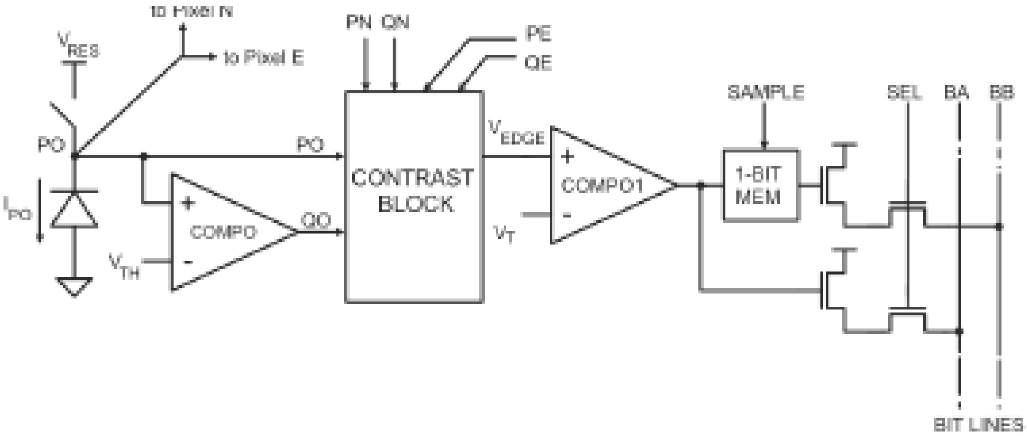


Figure 22. Schematic of the pixel-level contrast extraction circuit.

We consider three adjacent photodiodes characterized by a different light value: PO is the less illuminated pixel and PN is the most illuminated one. The contrast estimation process stats when the PN exceeds the threshold Vth1(ON=0) and stops at the same threshold is across by the less PO illuminated pixel.

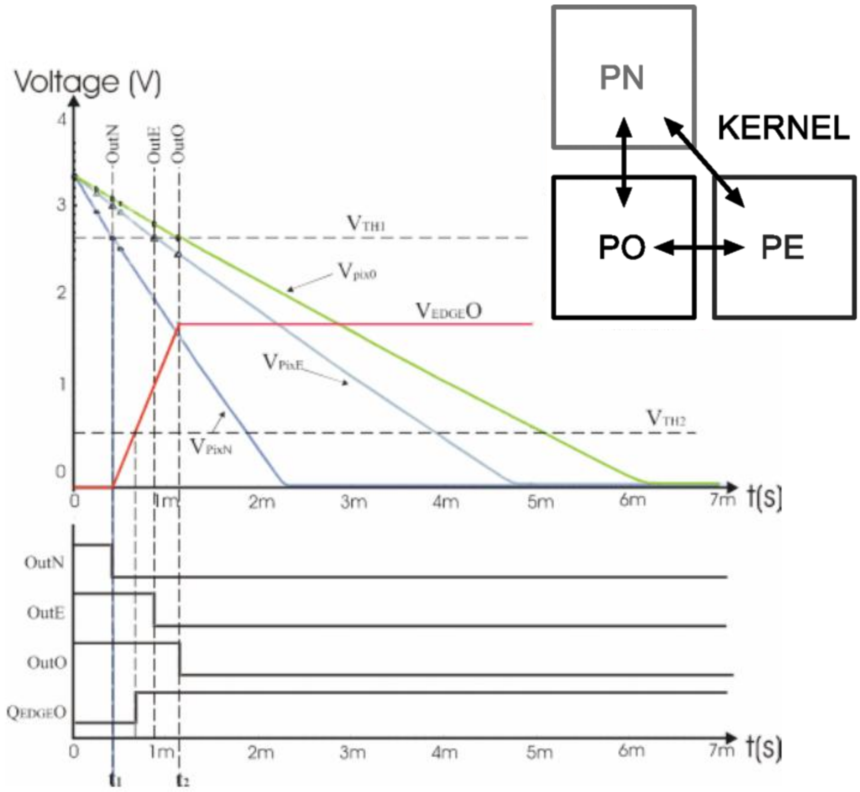


Figure 23. Shows a timing diagram of a frame acquisition.

After the reset phase any pixels star a voltage discharger ramp due to the incident light. The Vpix0 is connected to contrast block with other two analog signal VPN VPE and the other three comparator (OO,ON,OE).

During this contrast estimation process the Vc is sampled and quantities regulated by this formula $VEDGE0(t2) = VPO(t2) - VPN(t2)$.

The resulting normalized contrast is then binarized by means of comparator Comp2, it compared contrast to Vth2 set by user. The output can be stored into a 1-bit memory (Sample) or directly provided on one of the two bit-lines of the pixel as current information.

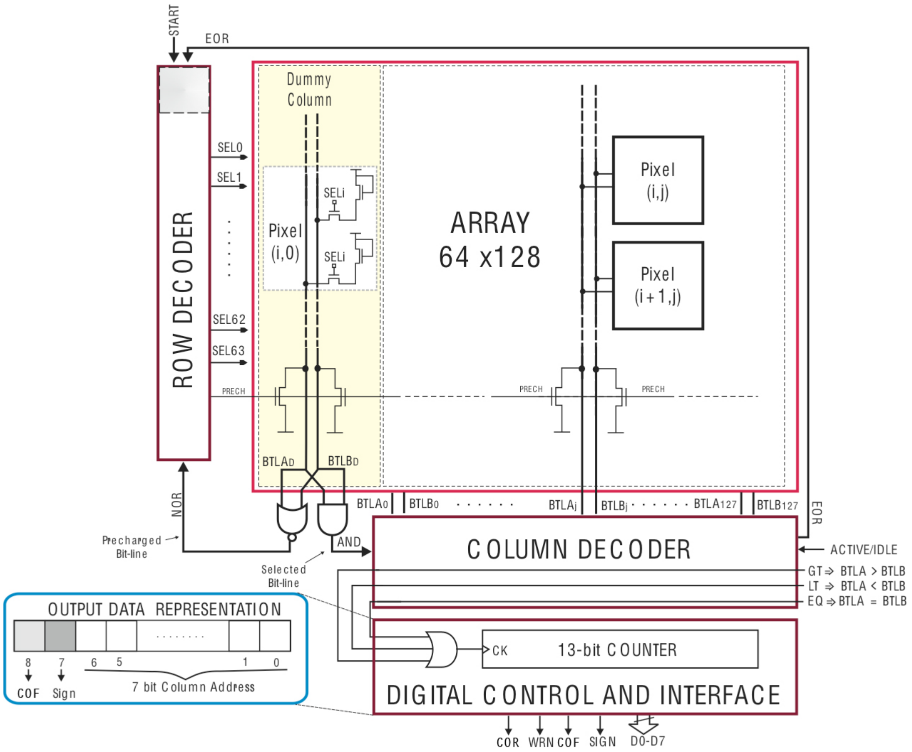


Figure 24. Block Diagram of the Spatial contrast sensor.

The internal architecture of the sensor, visible in (Figure 24.), comprises in addition to the matrix of pixels, also the logic required for the management of the array and output data; it has the aim to minimize the amount of control logic from added externally.

The Row-decoder select the 64 lines of the image sequence. The pixels of the row selected write two bits on the respective lines: BTLA for the current frame BTLB for the previous frame.

The Column-decoder controls the possible disparity between BTLA and BTLB in sequence and provides a pulse on one of the three bit lines (GT, EQ, LT) depending on the type of disparity, after which increments the counter.

Only the disparity presence, the column address to 7-bits of the relevant pixel is put out to the chip with its sign bit (SIGN). After the last address of the same row, the bit

COR is made logic high represents a newline, and passes the control to the next line. At the end of the process of image reading, then after 64 pulses of the COR, the bit of the end frames (COF) assumes high logic value and the sensor is stopped a new command are waiting to start another cycle of acquisition. The column address of the current pixel and placed at the exit asynchronously in blocks of data at 80 MB/s.

The *counter* to 13bit about the digital interface it's used in different ways on the different mode closed, Active or Idle.

In *Active Mode* the sum of the pulses on three bit lines is always 128 number per Frame. So, putting the three-bit-lines as input to the counter clock signal, its value will be equivalent to the address column of pixels processed. When a disparity is detected, the counter value, which corresponds to the column of the pixel related, is incremented and sent out together with the sign of the disparity. Differently, if any disparities are detected, the counter is still incremented, but the data is retained, maintaining the output data very low. At each end of the row the counter is reset and the process for the Next line begins.

In *Idle Mode* only on the pulse of the column decoder lines that indicate a disparity (GT and LT) are considered without resetting the counter at each end of row, at the end of the scanning process of the frame the in a counter the total number of disparities present in the image will contained. During the entire raster scan of the sensor any data at output are delivered, in fact, only at the end of the acquisition phase the value stored in counter 13 bits will be delivered. The data will be divided into the upper part and lower part due to the limited number of lines.

The data flow bandwidth is organized whit a 10-bit incremental code (Figure 25.). The 7 right-most bits identify the pixel column address (128 pixel/row), D7 is the sign of the gradient (it is only used in Motion Extraction), COR detects that next data will belong to the following row of the imager COF identifies the end of frame.

D0-D7 are synchronized with WRN and can be directly read out from the counter. COR and COF are set with asynchronously. Data flow organization permit to elaborate directly the position information of the active pixels, so to permit to reduce computation and bandwidth.

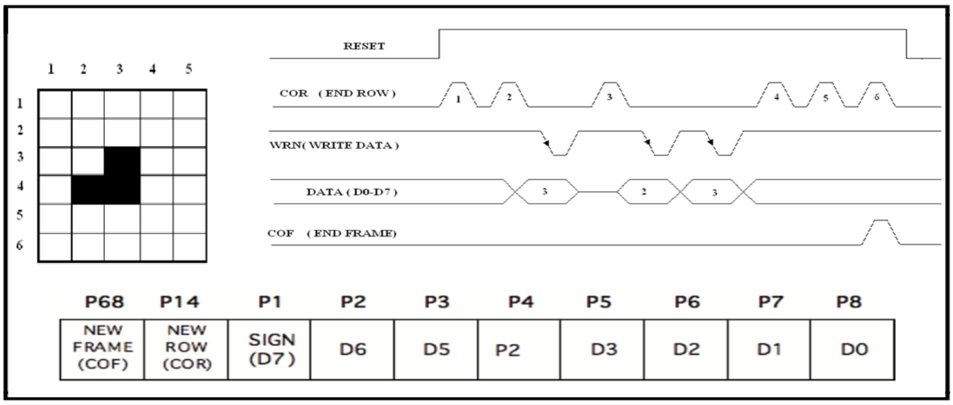


Figure 25. Example of data coding and temporal input output signalas.

Now Considering one row of the array, for each pixel the address is delivered according with the raster-scan mode.

The Figure 26. Represent a image portion in active mode extracted an corresponding code.

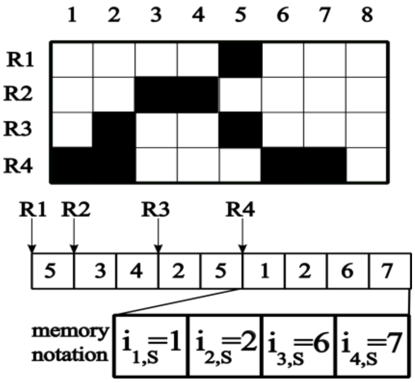


Figure 26. Portion image, data coding and memory rappresentiation.

If we consider the pixels within row R3 and R4, although pixels 2 and 5 occupy contiguous date output, they are physically placed at a distance of 3 pixels from each other in the array.

But main issues here is that binary contrast is in fact a fairly poor information, which is not reliable enough for certain applications. But vision sensor are the ultra low power characteristics in fact draws approximately 100uW at 3.3V at a frame of 50fps in a sparse-matrix through a positional coding.

The unconventional positional data coding permit to reduce the bandwidth on the other hand open problem to maintain the same efficiency on algorithm elaboration.

3.2. Temporal contrast vision sensor

Differently the spatial contras sensors the temporal contrast sensors are based on pixel to pixel comparison between two different frame over the time. In this way only the motion events caused by the moving are delivered and automatically remove the static background.

The sensor during the integration time acquire a intensity value after that compare this value with two memory value simultaneously (SIMD). Where in this memory are keep the low values and the high reelected of low and high acquire during the previous frames.

Most of these implementation, are very sensitive to the threshold Th and works correctly only in particular conditions of object speed and frame rate.

The advantages of this sensor is the possibility to compare the gradient of two frame consequently (F_i-F_{i-1}) with a dynamic threshold.

In fact the threshold of sensor Th is modeled as an exponential moving average:

$$Th_{j+1} = \alpha \cdot F_j + (1-\alpha) \cdot Th_j$$

where F$_j$ represent the current frame and α a constant smoothing factor between 0 and 1. This peculiarity permit to adapt the sensibility of movement through the α parameter. With the aim to obtain dynamic thresholds the sensor using a low pass filter.

This architecture should be implemented pixel by pixel and it would be impossible the CMOS integration. For this reason it was necessary to find an alternative method for large scale hardware implementation.

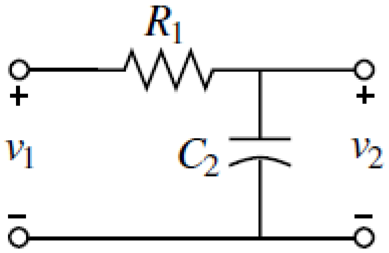


Figure 27. Conventional low pass filter.

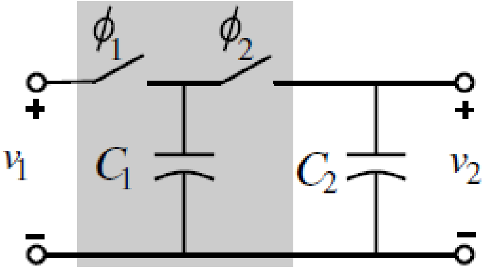


Figure 25. Equivalent low pass filter with switched capacitor.

The method adopted is based on the equitant between the rc network and the switched capacitor.

This circuit was developed by replacing the resistor, *R*$_1$, of the standard low pass filter circuit with the parallel switched capacitor resistor circuit opportune controlled.

$$R_1 = \frac{1}{C \cdot f_{ck}}$$

The value of this resistor decreases with increasing switching frequency at the same time Compatibility with CMOS technology is obtained.

In order to understand better de characteristics of the sensor the Figure 25. Represent the hardware implementation on chip.

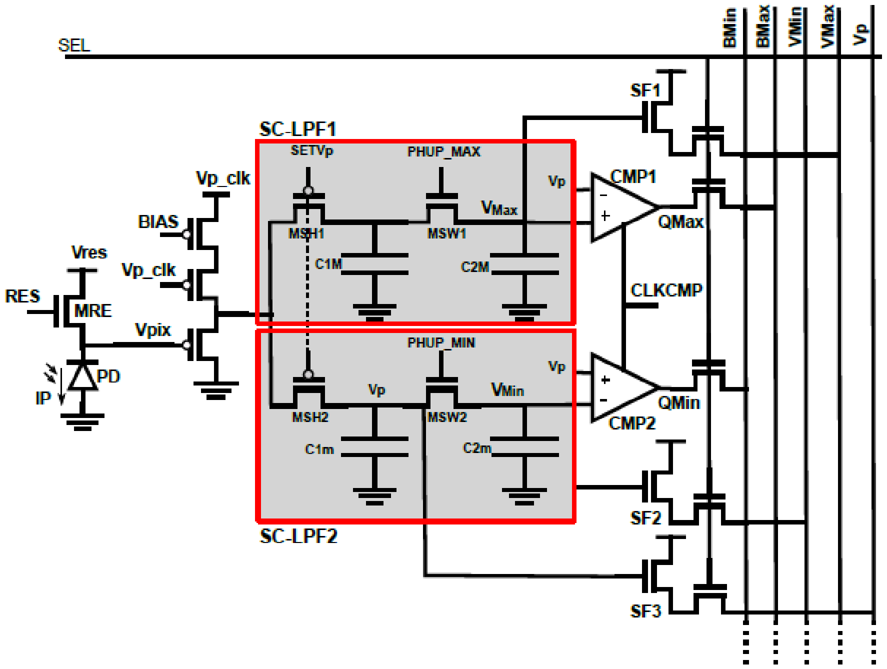


Figure 28. Pixel schematic of temporal contrast sensor.

The photodiode (PD) of Figure 28. works in storage mode, with a source-follower readout transistor, which is turned on by Vp clk only when necessary, reducing the pixel DC power consumption.

The two SC-LPF1/2 starting from the value VP proportional at the photodiode current extracted, evaluate the VMax and VMin values respectively for any frame rate.

In the first step the Current Vp extract by the photodiode is sampled trough the SetVp input on temporal memory C1M and C1m capacitors.

The two output of the filters are stored onto the PMOS capacitor C2M (C2m).

In a second step trough a distribution when the switch PHUP is closed charge sharing takes place with this LPF transfer function:

$$H(s) = \frac{1}{1+s\tau_n} = \frac{1}{1+s\cdot\left(\frac{C2M}{C1M}\cdot\frac{n}{f_0}\right)}$$

where the C1M and C2M proportional, characterize the filter characteristics.

The two output of the filters are stored onto the PMOS capacitor C2M (C2m) in order to keep of past signal variations for the next comparison .

At the end of the integration time, the CLKCOMP activated the comparators (CMP1, CMP2) and compare VP with VMax and VMin respectively stored in two analog memories, generating the two bits QMax and QMin.

Differently as previous sensor this sensors deliver all pixel value in a raster scan mode, but the binary form it’s common.

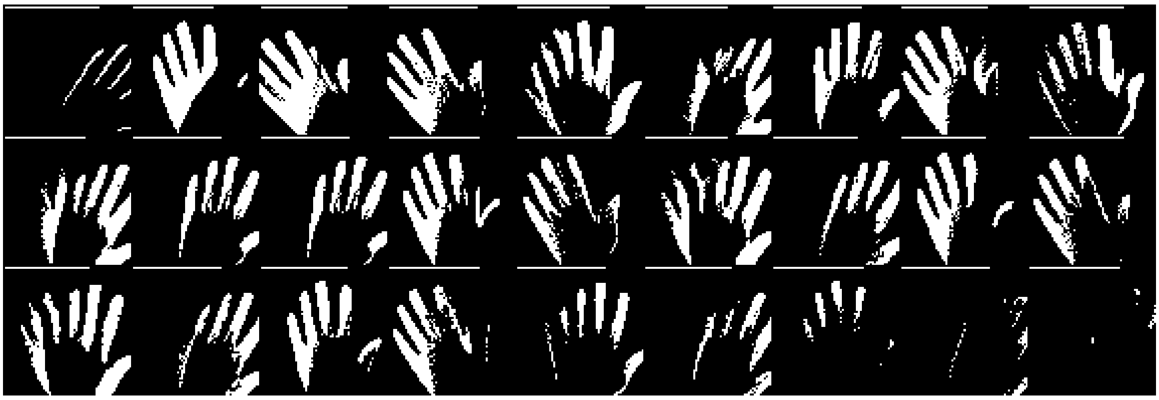


Figure 29. Temporal sequence of object absorption.

The output binary signal (QMax, QMin) delivered by each pixel are collected at the array level and can be processed outside the chip by the higher-layer algorithm trough different n value setting, implementing high level vision tasks.

In fact, the high level algorithm will be set different n value, where n > 0 represent the rapidity event absorption, fastest response (n=1) each frame the memory are updated, events are not suppress (n→∞) the memory are never updated .

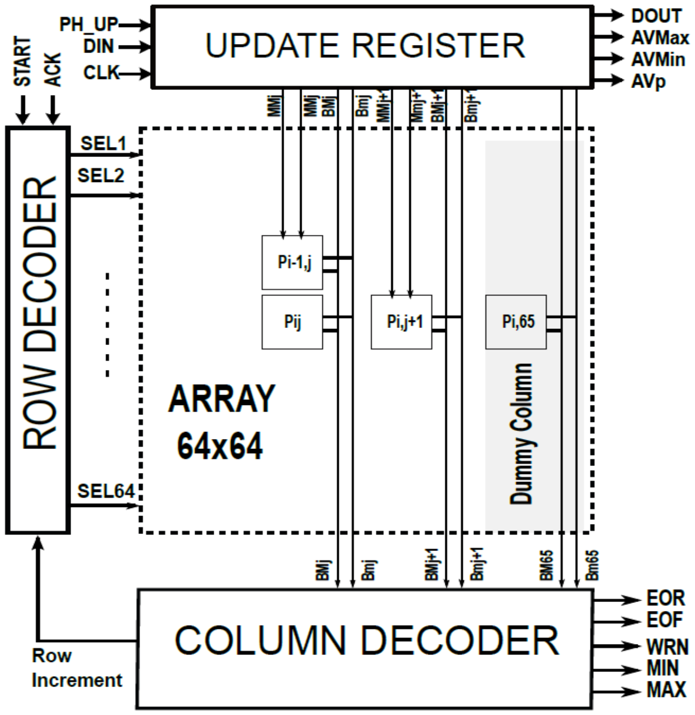


Figure 30. Block diagram of temporal difference sensor.

The Figure 30. shows the block diagram of the sensor architecture for an array of 64x64 pixels prototype. The imager is an addressable array of pixels, with a 64-cells ROW DECODER and a 64-cells COLUMN DECODER. The UPDATE REGISTER consists of a 64x2-stages shift-register with two main functions:

READOUT: after a row-selection, bit-lines are loaded into the UPDATE REGISTER and read out serially, through DOUT, CLK;

UPDATE: after a row selection, a 64x2-bits binary mask is serially loaded into the UPDATE REGISTER, through DIN, CLK. PH_UP is pulsed, updating only the selected pixels of the row (MMj, Mmj). Next row is selected and a new row of masks is loaded. This characteristic permit to tune the response filter dynamically according whit the external algorithm setting, it generate a cooperation between sensors and external digital processing.

Chapter 4

4. Custom stereo vision system

We will concentrate our work on of custom sensors and custom algorithm optimization within the optimized embedded system realization. In the first part of this chapter we will analyze the principle of stereo vision and evaluate the possibility of applying this technique to a custom vision sensor with positional data coding characteristics.

The sensor has evident low power characteristics, in addition to the efficient data representation that exploits on the one side the advantages for compression and on the other side a new algorithm for elaboration phase are required.

The positional data coding eliminates the redundant information present in the image, this type of coding has all the characteristics to be exploited for images comparison. As we can see in this chapter the stereo vision is a typically complex computational technique based on the matching of portions image extracted for two different point of view. Stereo Vision System is a leader much studied considering the numerous applications in both the private and industrial, especially in the latter where applications often require three-dimensional passive monitoring devices.

SV has been at length investigated and a large number of algorithms have been developed for its computation. A general overview of stereo vision algorithms is accessible in [44] based on standard data. There are currently many important stereoscopic products capable of providing both synchronized cameras and the stereo matching software. Only few of them take into consideration the power consumption or the hardwire connection.

In fact, these systems are targeted to applications where high resolution and color information are of main concern.

Stereo-vision algorithms require a highly intensive signal processing based on spatial matching with large operation redundancy. In this context, the use of a custom vision sensor could have the advantage of pre-selecting the data that will take part in the processing, thus minimizing the subsequent amount of operations [18].

4.1. Custom Stereo algorithm

If we analyze a single image is not possible to reconstruct the three dimensional structure of the observed scene. This is due to the loss of information in the perspective projection, which maps points in 3D space in a 2D space.

Like the human visual system, you can place two cameras at a certain distance from each other and receive an image from two slightly different points view.

In humans, these distinct images are used to estimate the depth and fuse together to create a single image of the scene.

The estimation of disparities is the problem of finding corresponding points in a pair of stereo images to calculate the distance of the object. Il literature are large number of stereo algorithms are presented, but only a few are tailored for custom sensors [45][46]. In fact, most of the activity on image processing, developed by the scientific community, has been based and tailored on standard imagers. It s therefore difficult to share and to exploit this know-how on a custom sensor architecture.

In order to develop a new algorithm is necessary to consider the stereo system geometry. To analyze the geometric relationships that between three-dimensional coordinates of a point of a scene and the coordinates of its projection on the image plane, a model based on an ideal optical camera is used. This model does not include any distortion due to the lenses. Moreover, the image plane is considered to be continuous, while many current sensors, being composed of cells, have, in fact, only quantized coordinates. Given the simplifications assumed, it is considered an ideal model of camera; it is representative and useful because it allows us to focus on complex geometry, avoiding the complications due to the complex optical geometry of the real objectives and inevitable spurious factors, including distortions and aberrations, which occur in practice. Imagine placing two cameras on the same x axis (symmetrical points in respect to y axis) positioned in the same direction parallel to y and lying on the z = 0 Figure 31.

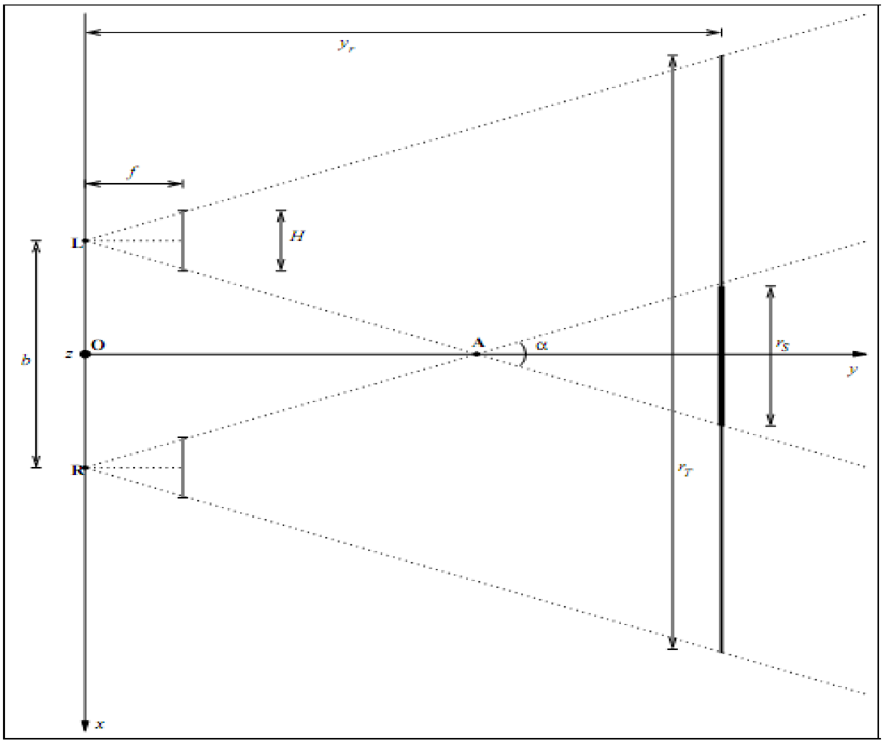


Figure 31. Top view of the stereo geometry.

The cameras have slightly different points view caused by the distance b between sensors Figure 32.

This baseline can effect on their respective image plane named disparity. It’s possible to understand the object distance trough the disparity exploitation.

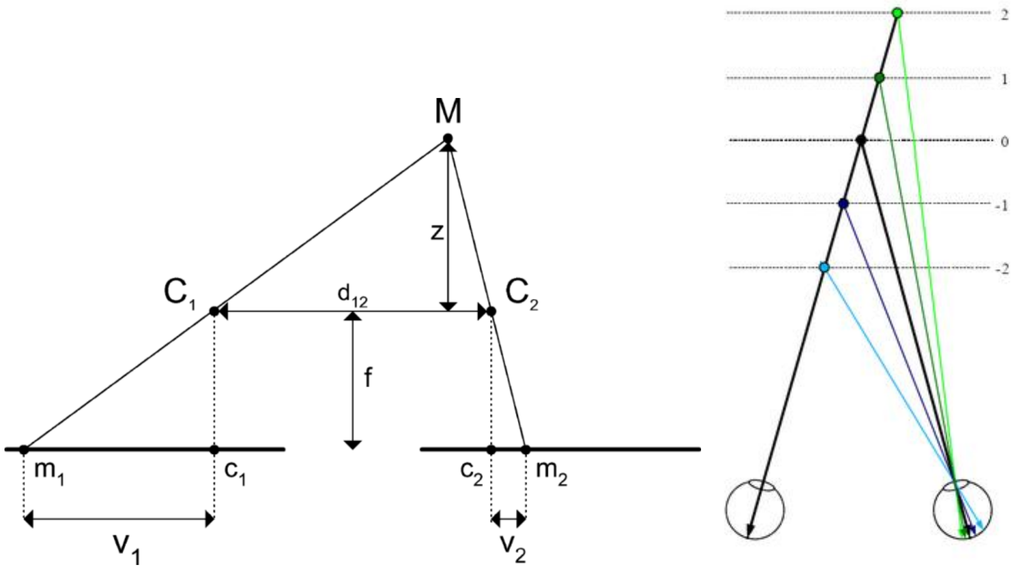


Figure 32. Top vision from the pattern of cameras and differente disparity proiection.

Now the stereo probleme it’s a correspondence problem, can be solved using an algorithm that scans both the left and right images for matching image features.

In literature several algorithms exist and they can be divided in two groups:

- *Correlation based*
- *Feature based*

In the *Correlation based* algorithm it’s possible adopt local or global methods, the disparity is evaluated using a winner-take-all (WTA) strategy.

In a local methods the disparity of each pixel is calculated without considering disparity computed of other pixels. This correspondence can be ambiguous. For this reason the image for matching is subdivide in windows (5x5 pixel or 7x7 pixel i.e). The blocks of a image are compared with those of the image to search for correspondence. Between two blocks the similarity can be very well measured by calculating: Normalized Cross Correlation (NCC), Sum of Squared Difference (SSD) or Sum of Absolute Difference (SAD).

- Normalized correlation: $\frac{\sum\sum L(x,y)\cdot R(x,y)}{\sqrt{\left(\sum\sum L(x,y)^2\right)\cdot\left(\sum\sum R(x,y)^2\right)}}$
- Sum of squared differences: $\sum\sum\left(L(x,y)-R(x,y-d)\right)^2$
- Sum of absolute differences: $\sum\sum\left|\left(L(x,y)-R(x,y-d)\right)\right|$

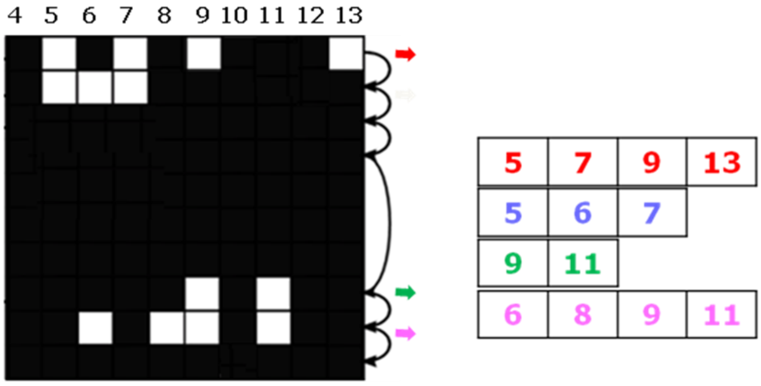
Moreover, the windows size is a trade-off because in a small area images are more similar despite of different view points, but in big areas, the ratio of signal to noise increases. These algorithms are suitable to treat a wide variety of images and provide a dense depth maps.

The *feature-based* algorithms face the problem of correspondence at a higher level than correlation-based. The first step is to identify items or groups of points in the stereo pair with certain features: edges, lines or angles. The differences between feature-based and correlation-based algorithms consist in the fact that the first exploit additional information in order to apply the matching, for instance the orientation or length of the edges. This is the main limit, because it is not always possible to know the type of feature a before as it depends on the applications.

Our kind of images does not allow the use of classical philosophy to calculate the disparity because our sensor does not provide conventional images as seen above.

For these reason is necessary to develop a new custom algorithm able to exploit the peculiarity of sensor efficiently.

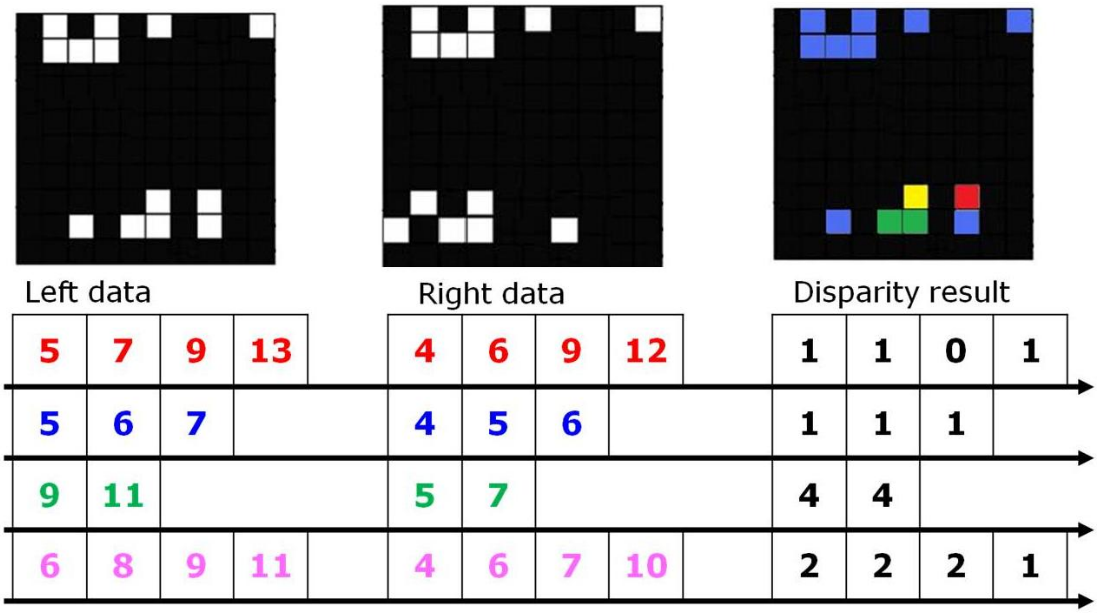
The sensor extract the spatio-temporal contrast and binarized these information delivering only the address related to active pixels.



5	7	9	13
5	6	7	
9	11		
6	8	9	11

Figure 33. Sensor image date deliver in spars matrices organization.

This sensor execute a pre elaboration of the image extracting the edge present in the scene directly on chip and deliver a spar matrices organized in this way. In this scene are presented 13 pixels and only the corresponds address are delivered in a raster scan order. Figure 33.



Left data				Right data				Disparity result			
5	7	9	13	4	6	9	12	1	1	0	1
5	6	7		4	5	6		1	1	1	
9	11			5	7			4	4		
6	8	9	11	4	6	7	10	2	2	2	1

Figure 34. The pair of images left and right respectively e the ideal result the disparity algorithm are represented.

Figure 34. Figure 35. describes the problem to be resolved, in the top part is possible to see the left and right images delivered by the sensors, on the right there is the image we wont to obtain by algorithm.

In this last image every pixel is coded with the a color which represent the depth of that pixel.

Below the previous images the corresponding data coding is shown.

The task is to takes advantage from the characteristics of the sensor evaluate the disparity of object present in each scene, starting on this positional data coding.

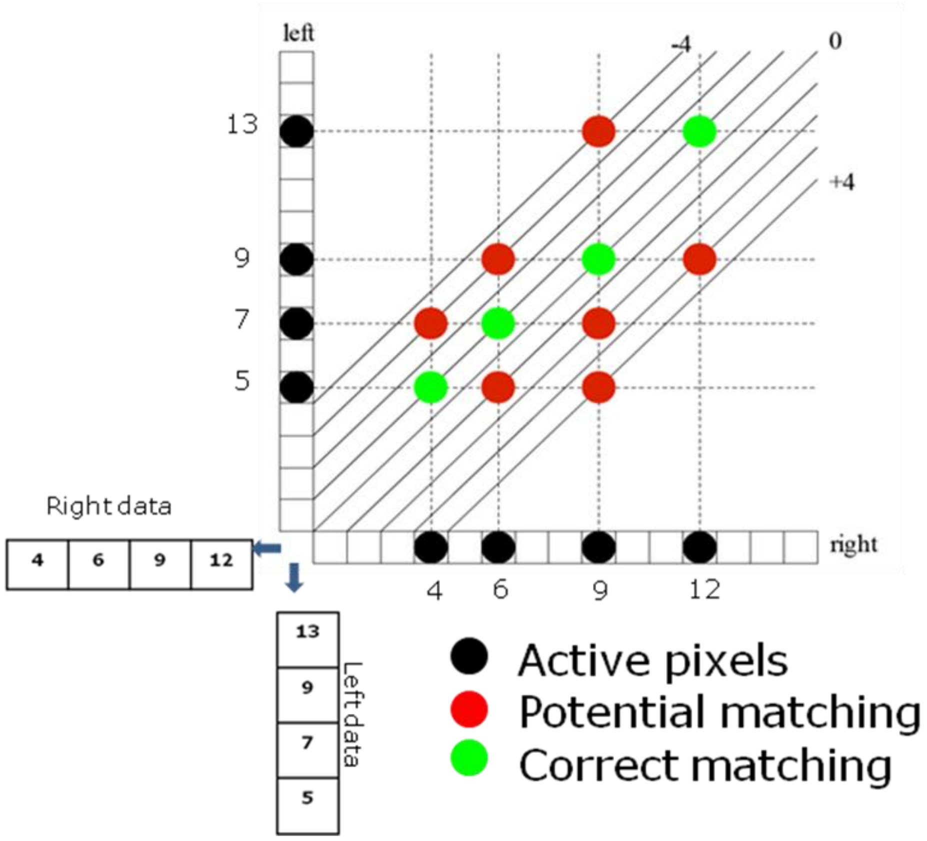


Figure 35. In vertical line we represented a row of left image end in horizontal we represented the same row delivered by the right sensor.

In order to explain the objective of this algorithm we represent the possible matches in Figure 35. , where in vertical line we have represented a row of left image end in horizontal we have represented the same row delivered by the right sensor.

Looking at this representation is necessary to define some constraints:

- Epipolar (The disparity is evaluated only row by row)
- Uniqueness (It's possible to find only one correct match between left end right rows)
- Ordered (The sensor deliver the address in ascending order)
- Disparity limit (Knowing the geometry of the system is possibility to limit the maximum disparity evaluated).

Bleak point represents the active pixel in the same row, red represents the possible candidate matching and in green the correct match. And the diagonal line represent the same disparity information for any mach associated.

The idea looking at the data is to adopt the Graph Theory [47][48][49]. This type of approach is a mathematical structures used to model pair wise relations between objects from a certain collection, in this case the addresses of the pixels from left and right imager.

In the theory graph the interconnected objects are represented by mathematical abstractions called *vertices(green)*, and the links that connect some pairs of vertices are called *edges(blu).*see in Figure 36.

It's interest to note for any correct *vertices* the connection must move up and to right.

These movement are related on the disparity point skipped in scanline and the disparity mismatch ripest the previous match.

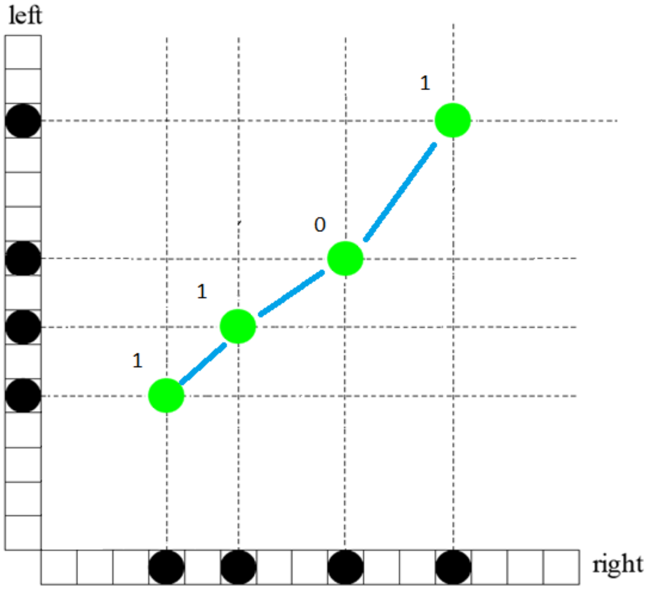


Figure 36. Two corrispondent row and the disparity.

The goal of the function cost is to balance the point skipped and the disparity related associating the minor cost to correct match Dijkstra inspired [58].

In order to do this, the function cost is to builder including three principal element, the first and second element evaluate the point skipped in left and right scanline. In other word the non adjacent address value are penalized and the last term the variation of disparity respect the last match found, normalized with the distance between the adjacent pixels.

We consider this correct sequence matching:

$$S_{\mathrm{j,p}} = \left(i_{k,l}, i_{q,r}\right)\ldots\left(i_{j,l}, i_{p,r}\right) \quad \text{for } k,j,p,q \in \{1,2,\ldots,\mathrm{M}\} \tag{1}$$

The function cost associated is:

$$\begin{array}{ccc} \mathrm{I} & \mathrm{II} & \mathrm{III} \end{array}$$

$$C\left(S_{j,p}\right) = \sum_{1}^{M-1} {}_{j,p} A\left(\left(i_{j,l} - i_{j-1,l} + 1\right) + \left(i_{p,r} - i_{p-1,r} + 1\right)\right) + f\left(d_{j-1,p-1}, d_{j,p}\right) \tag{2}$$

This type of function is additive and is based on the all legal previous match founded.

The idea is to build a cost function C(S) computed for any possible match between the corresponding rows of left and right imagers. In this way the stereo correspondence problem turns into a cost function minimization problem.

A represents the balancing coefficient which is used to calculate the cost function contribute of the new pixel added to a sequence, typically set to 0.5 (this parameter is connect to the distribution of the pixels). The last term on the rights the change in disparity, normalized with respect to the distance between the pixels in the left row, which can be expressed as:

$$f\left(d_{j-1,p-1},d_{j,p}\right)=\frac{\left|d_{j}-d_{j-1,p-1}\right|}{\left|l_{i_{j,l}}-l_{i_{j-1,p-1}}\right|}=\frac{\left|\left(l_{i_{j,l}}-r_{i_{p,l}}\right)-\left(l_{i_{j-1,l}}-r_{i_{p-1,l}}\right)\right|}{\left|l_{i_{j,l}}-l_{i_{j-1,l}}\right|} \tag{3}$$

In this case for any pare of two pixels delivered by the left and right sensors a function cost will be defined.

The cost function C(s) is additive. For a given $S_{j,p}$ ending in ($i_{j,l}$, $i_{p,r}$), the cost function C($S_{j+1,p+1}$) of the next sequence, adding the term ($i_{j+1,l}$,$i_{p+1,r}$), is obtained by simply adding the term:

$$A\left(\left(i_{j+1,l}-i_{j,l}+1\right)+\left(i_{p+1,r}-i_{p,r}+1\right)\right)+f\left(d_{j,p},d_{j+1,p+1}\right) \tag{4}$$

to the previous C($S_{j,p}$).

Until now we considered the ideal case where all points are perfectly matched. Dealing with real data, it is necessary to consider the occlusion problem, where some points in the scene are not visible by both sensors.In order to take into account this issue, the function cost C($S_{j,p}$) must be modified with respect to eq.(4):

$$\mathrm{C}\left(\mathrm{s}_{\mathrm{j,p}}\right)=\mathrm{C}\left(\mathrm{s}_{\mathrm{j\text{-}v,p\text{-}w}}\right)+A\left(\left(i_{j,l}-i_{j-v,l}+1\right)+\left(i_{p,r}-i_{p-w,r}+1\right)\right)+f\left(d_{j-v,p-w},d_{j,p}\right) \tag{5}$$

Here, the next matching is not necessary referred to the next pixel located in the memory, but has to be found elsewhere inside the row.

Eq.(5) it is solved through dynamic programming; for any j and p pair, addressing the pixels in the memory, the cost function is computed with respect to all the pixels placed in the memory. Eq.(5) has a complexity M^4, where M is the number of pixels in the row of the sensor. This is a worst case estimation, due to four nested embedded loops connected to j, p, v, w parameter. However, under some assumptions the complexity of the problem can be significantly reduced without affecting the overall algorithm reliability. Rather than involving all the pixels of the row in the matching computation, only those located in the neighborhood can be involved by limiting the values of v and w in eq.(5).

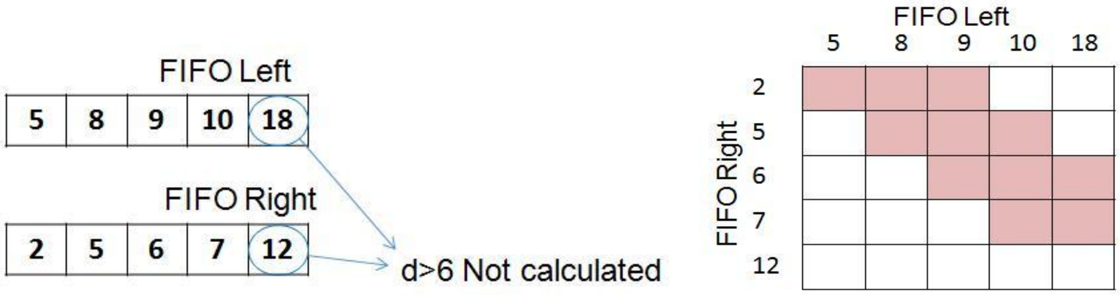


Figure 37. Example of cost function constraint.

Figure 37. shows an example of cost function calculation under the following assumptions:

- matching pixel neighborhood v, w = 3
- maximum admitted disparity is d_j=6 pixel. (this value depends on the specific geometry of the optical system);

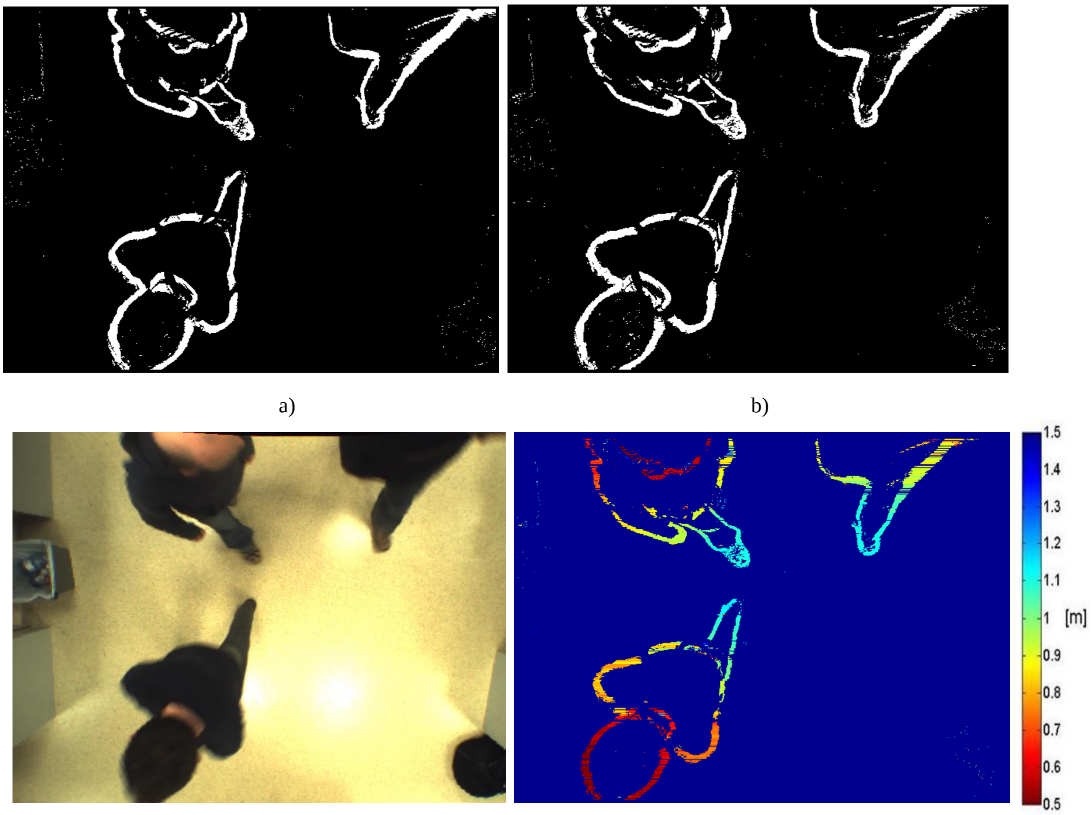
The cost function of the two rightmost pixels of the lines shown in Figure 37. (left) is not calculated, being their disparity larger than 6 pixels.

4.2. Stereo vision system simulation

After the implementation algorithm it’s necessary to validate the results applying the algorithm on a image characterized by the conventional resolution adopted for people monitoring. In order to quantitatively evaluate a disparity estimation approach develop a standard data-set with disparities is required.

First to test the stereo algorithm develop is necessary to simulate the data delivered by the sensor algorithm on stereo pedestrian data-set. Unfortunately, standard disparity data-sets, such as the *Tsukuba*, *Venus*, or *Map* data-sets may not be applicable for people monitoring algorithms. In fact these dataset include some ambiguous periodical pattern difficult discriminated only by binary edge detection. It’s necessary to adopt a data set rectified and synchronized (*Overhead* Scenario 2D pedestrian detection http://www.cdvp.dcu.ie/datasets/pedestrian_detection/).

The *Overhead* scenario is set in an indoor environment with the camera positioned at around 3 meters above the ground and orientated back towards the ground plane. The camera has a limited field of view and due to its proximity with the ground plane it does not encounter significant occlusion problems. The scene is brightly illuminated with a scene's lighting is stable. An example of two picture obtained by the sensors simulation are see in (Figure 38.). The disparity map obtained show the different part of the body clearly represented by different color, for example it’s possible to filter out the depth value correspondent to feet or the heat. In the disparity map (Figure 38. d) some non closing edge are presented due to non continuity border sensor’s extraction. In spite of de ambiguity of the binary images the algorithm is able to



a) b)

c) d)

Figure 38. Simulationa results a)Left binary image b)Right binary image c)Real scene d)Disparity resoults.

4.3. System specifications an realization

In order to develop a new stereo vision system in necessary to consider the geometry of stereo. To analyze the geometric relationships that between three-dimensional coordinates of a point of the a scene and the coordinates of its projection on the image plane, a model based on an ideal optical camera is used. This model does not include any distortion due to the lenses. Moreover, the image plane is considered to be continuous, while many current sensors, being composed of cells, have, in fact, only quantized coordinates.

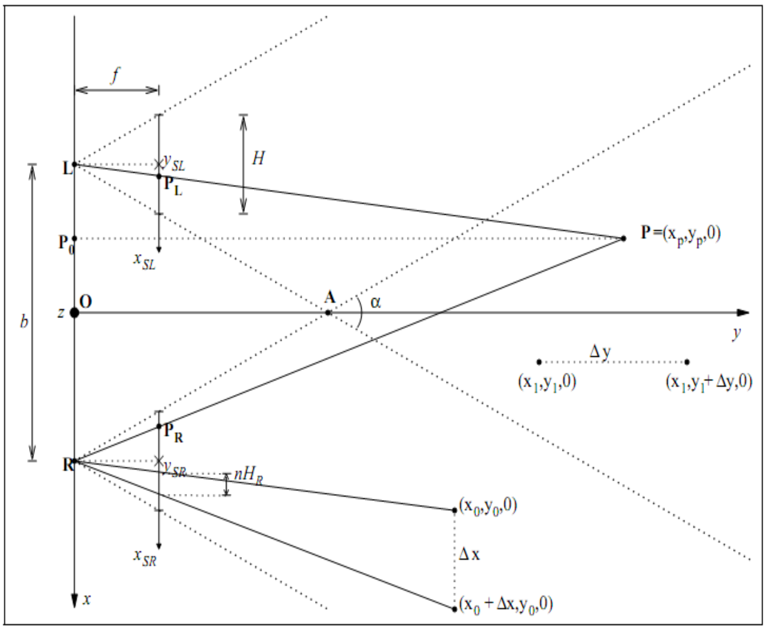


Figure 39. Top vision from the pattern of cameras.

The main element are:

- System coordinates;
  - x y z : coordinate word;
  - x_{SL} y_{SL} : coordinate left sensor;
  - x_{RL} y_{RL} : coordinate right sensor;
- System coordinates
  - f : focal length;
  - H : horizontal size of the sensors;
  - N_H: horizontal resolution (number of pixels);
  - N_R: horizontal size of a pixel $H_R = \frac{H}{N_H}$;
  - b : baseline represent distance between sensors.
- Other sizes
  - P : point in the real world coordinate (xp, yp, 0);
  - H : horizontal size of the sensors;
  - P_L P_R: projection of point P on the sensor on the left and right, respectively;
  - PL PR: projection of point P on the sensor on the left and right, respectively.

The P_L coordinates are (XpSL, 0) while those of P_R are (XpSR, 0).

The visible area from both cameras is bounded by:

$$\overline{OA} = \frac{fb}{H} \quad \alpha = 2arcan\frac{H}{2f} \tag{6}$$

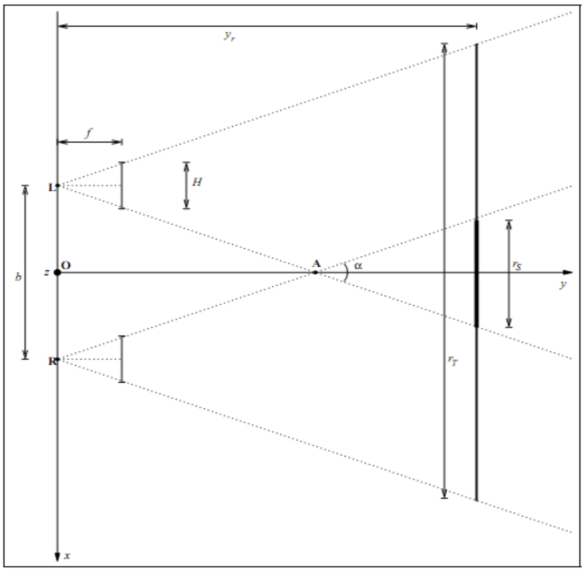


Figure 40. Top vision from the pattern of cameras.

With reference to Figure 40. Figure 39. we consider a parallel plan to XZ positioned with a distance Yr from the origin. The two cameras don't see the same part of that plan; if $y_r \geq \overline{OA}$ intersection between two images is null. In the bi-dimensional scheme the plan is represented by a straight line and the intersection of two visions by a segment of a length of:

$$r_S = \frac{Hy_r - fb}{f} \tag{7}$$

The union is instead represented by a segment of a length of:

$$r_T = \frac{Hy_r + fb}{f} \tag{8}$$

We define the overlapping degree r $\in$(0,1) the relation between intersection and union of the two visions.

$$r = \frac{r_S}{r_T} = \frac{Hy_r - fb}{Hy_r + fb} \tag{9}$$

The degree of overlapping does not depend only on the parameters of the system, but also on the distance yr. Assuming that the algorithm for stereo analysis has particular requirements r>r$_{rmin}$ we have an estimate on the minimum distance:

$$y_r \geq \frac{1+r_{\min}}{1-r_{\min}}\frac{fb}{H} \tag{10}$$

These equations can be used to select the cameras more appropriate for a certain application. The angle α, the distance minimum and the maximum of the objects can be estimated from parameters of the cameras and the appropriate assumptions

- t : minimum measurable disparity from stereo algorithms;
- n : size of the more discoverable small object measured in pixels;
- Δ_x : size of small object that you want to detect;
- a : ratio of maximum permissible uncertainty on the distance of an object and its actual distance;
- R$_{min}$: minimum degree of overlapping for stereo analysis.

$$y_{\min} = \min\left\{\frac{a}{1-a}\frac{fb}{tH_R}, \frac{f\Delta x}{nH_R}\right\} \tag{11}$$

$$y_{\max} = \max\left\{\frac{fb}{H}, \frac{1+r_{\min}}{1-r_{\min}}\frac{fb}{H}\right\} = \frac{1+r_{\min}}{1-r_{\min}}\frac{fb}{H} \tag{12}$$

After these mathematical considerations it is necessary to set the constraints for a particular application and to adopt a stereo algorithm for obtaining disparity maps.

For our people counting applications we position the system at a height of 3m and see 1 to 2 meters away.

We fix: t = 1, a = 0,25 % , r = 85% , b = 8 Cm.

The next step is to confirm the theoretical calculations made by laboratory tests.

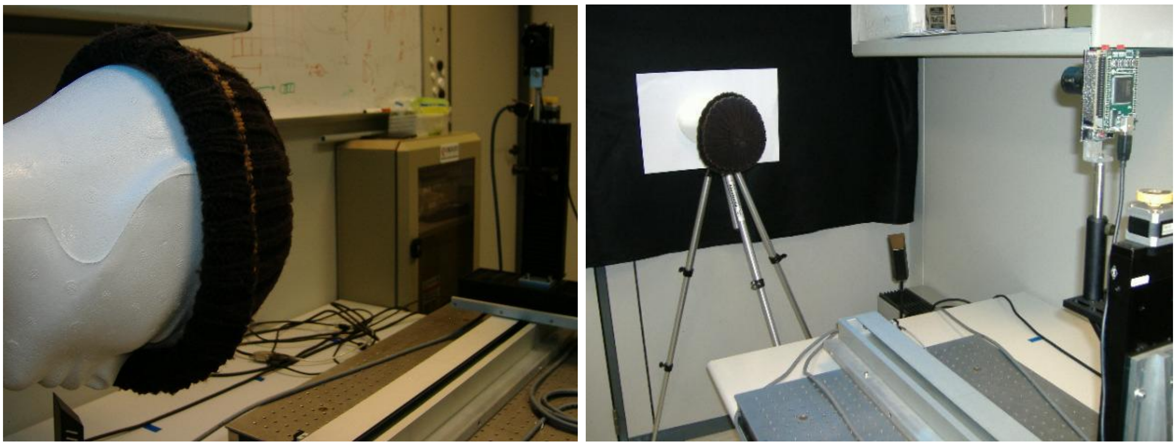


Figure 41. Laboratory Stereo vision simulation.

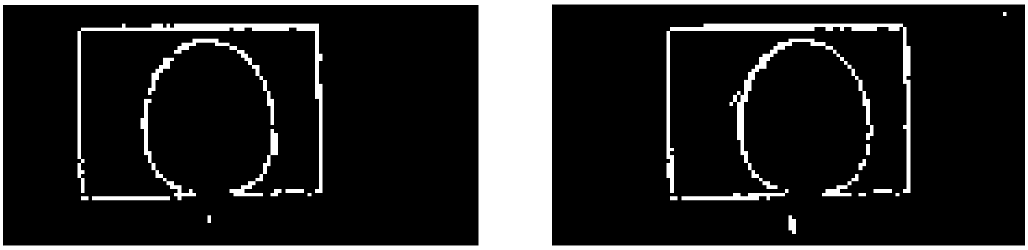


Figure 42. This picture represents left and right vision extracted by the same camera.

Figure 42. shows the setup realized in the laboratory to simulate the stereo visions. Only one sensor was mounted on a triple axes rail, allowing it to be shifted along the axes: Y (lateral), Z (depth). By using this system, the following results have been obtained:

Real distance (D_R): D_R=0,6m (with 6mm optics),

Estimated distance (D): disparity 28 pixel(shoulders), distance 0,65m, error 0,05 ≈ 8%

As we can see in the image there is a difference between the two shots of about 28 pixels.

Through geometric calculations, are able to calculate the distance from the system:

$$D = \frac{b \times f}{d \times H_r} \tag{13}$$

This error is caused by the low resolution of the sensor, using two cameras this error will be larger because it introduces new parameters that characterize the diversity of the two sensors as different focal lengths, misalignment, etc., these problems are solvable with the calibration.

To realize the first prototype tests, a board was designed through the Orcad software exploitation. The aim of the board is to create, starting from two sensor a single device easily connectable with a single FPGA. The second sensors were placed on the board at a distance regulated by the epipolar geometry, in particular controlled by the formulas (11) (12) illustrated above, fixing some parameters related a type of approximately 1.5m wide and 3m fixing in height.

The baseline was fixed at 8 cm, using a varifocal lens from 2.9 to 8.2mm, whit a zoom that could adapt the geometry for different situation. On the board mounted point where installed to allow the alignment of the sensors with the optics.

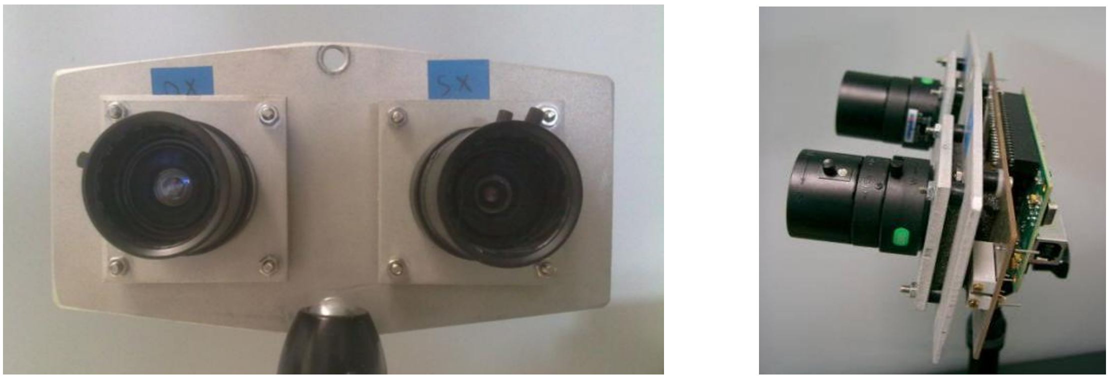
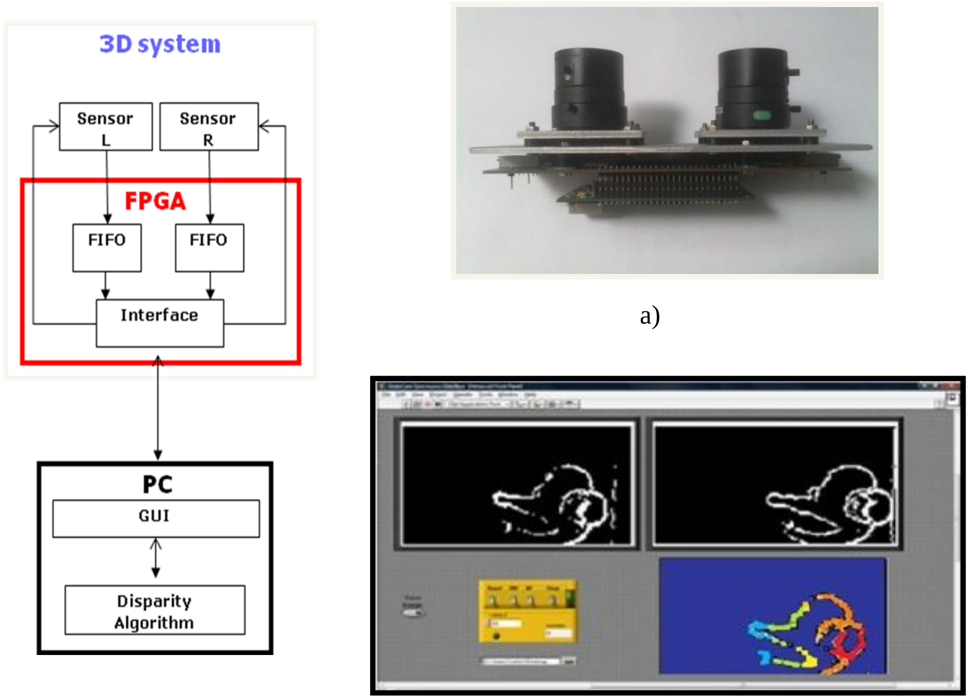


Figure 43. First Stereo prototype

We went on to FPGA programming phase using ISE software, supplied by the producer, using the VHDL language some components was realized for the correct management of control signals for two sensor and the data output.

The FPGA, Xilinx Spartan-3 has been selected for the development phase. In particular, XEM3001 Module [51], has be used, offering sufficient flexibility and complexity for the specific application.



a)

b)

c)

Figure 44. a) First prototype, b) Block diagram, c) GUI.

Data delivered by the sensors are delivered asynchronously by the sensors. Therefore, they need to be stored into a suitable to be two 8-bits FIFO memories built in the FPGA, For each pixel address, 7-bits are devoted to code the pixel position, while the MSB is dedicated to the End Of Row (EOR) signal. In fact, the last active-pixel of the row, has its MSB=1. This means that next data belongs to the next row of the imager. After reading 64 EOR, the entire image has been read out.

The only parameter, that can be changed by the user is the integration time, and can be tuned at the GUI interface, which is realized in Labview. The graphic interface also allows to display the binary images of the two sensors together with the colored disparity image and to store sequence of images on the pc, allowing the debug of the system.

Before the installation, we have tried to understand the distortions caused by the non idealities of the system. In general these distortions are corrected by software calibration.

The goal of calibration software is to determine two sets of parameters, intrinsic and extrinsic, in order to compensate the difference between both cameras. Infect, even if the camera are quite similar, optics introduce some distortions as different optics lengths and axis misalignment. The intrinsic parameters is used to correct the distortion of the lens and the difference in focal length, while the extrinsic space determine the offset of the two cameras, including the distance between them and the deviation from the parallelism of optical axes. Through these parameters it is possible to transform captured images in ideal pictures, as they would be seen by those pinhole cameras with parallel optical axes. This operation is a off-line procedure. For more details the reader can refer to [52]. In the literature there are several techniques for the calibration of a stereoscopic system, these techniques are based on geometric patterns, whose characteristics size and position of features are exactly known. Through the acquisition of a pattern such as a chessboard in different positions Figure 45. and then the user, looked at the images and fixed the intersection of the different corners of the pattern [53] the parameters, intrinsic and extrinsic, of the stereoscopic system are estimated.

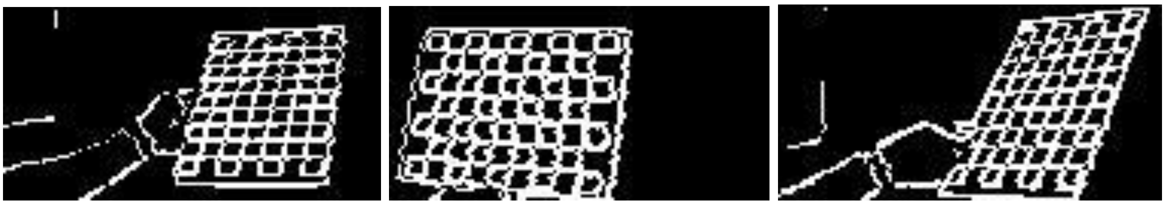


Figure 45. Calibration scene

As we can see in the Figure 46. after the corner extraction, the output of the calibrations is automatically generated. The output is a representation of the different projection of the scene acquired a file with the distortion values and the projection of error.

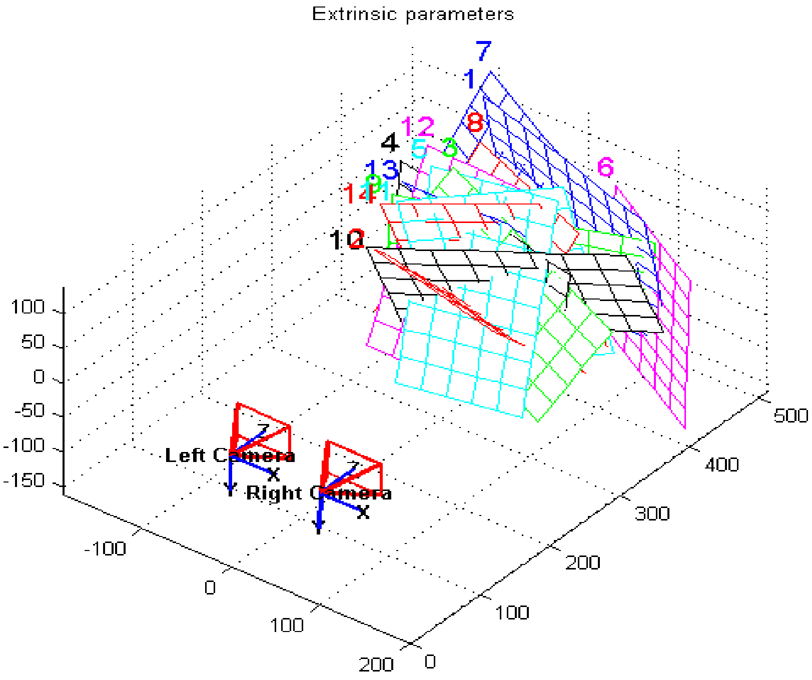


Figure 46. Different projection of the Calibration process.

After the calculation of the intrinsic and extrinsic parameter the ratification transform the two 2D image to establish a correlation.

Because of the poor resolution of the sensor, this operation wasn't very easy nor reliable; the parameters extracted were corrupted by major errors. See picture Figure 47.

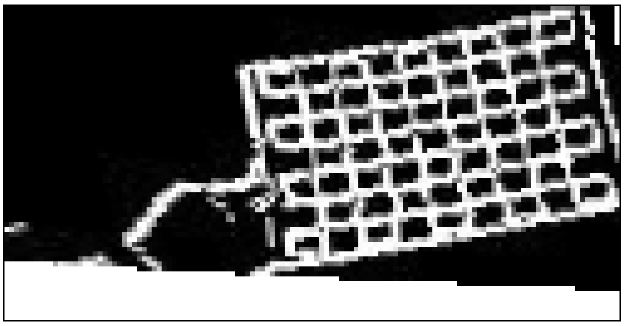


Figure 47. Rectifications resolut.

Typically the ratification is carry out on standard images, but applying this technique to the image produced by the sensors adopted, you lose the single correspondence between address and pixel. If we had applied this technique to the image, we would have obtained a sub pixel resolution, and grey scale.

This would have made it impossible to implement the hardware to calculate the disparity using an approach similar to that of the sensor proposed by Philipp [54]Figure 36. , where the disparity is calculated by direct comparison between the pixels using a Loser Take All circuit.

4.4. Experimental results

After the first demo realization the algorithm was tested using real date delivered by the vision sensor placed at the ceiling of a corridor and monitoring people walking through.

The two sensors are mounted with a baseline b=8 cm, in order to tune the system at a distance range between 1.5m and 3m. By using an f=7mm objective, the aperture reaches 20˚ along the x axis and 10˚ along y, having the sensor an aspect ratio of 2:1. In this phase of the work, many simplifications have been made, assuming the Moreover, in order to avoid problems coming from sensors misalignments, a manual mechanical alignment has been applied to the system, assuming both sensors to be properly placed on the *x* axis (symmetrically with respect to *y* axis) and with their focal planes orthogonal to *z*, as shown in Figure 49.

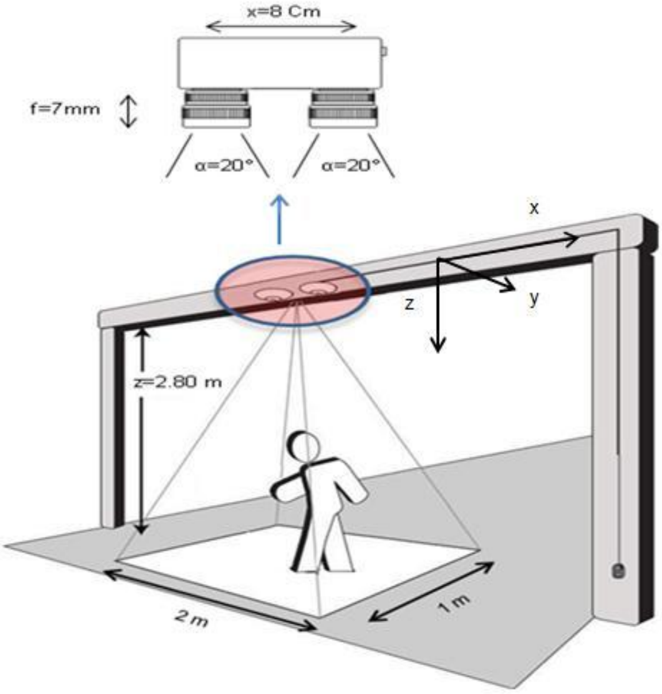
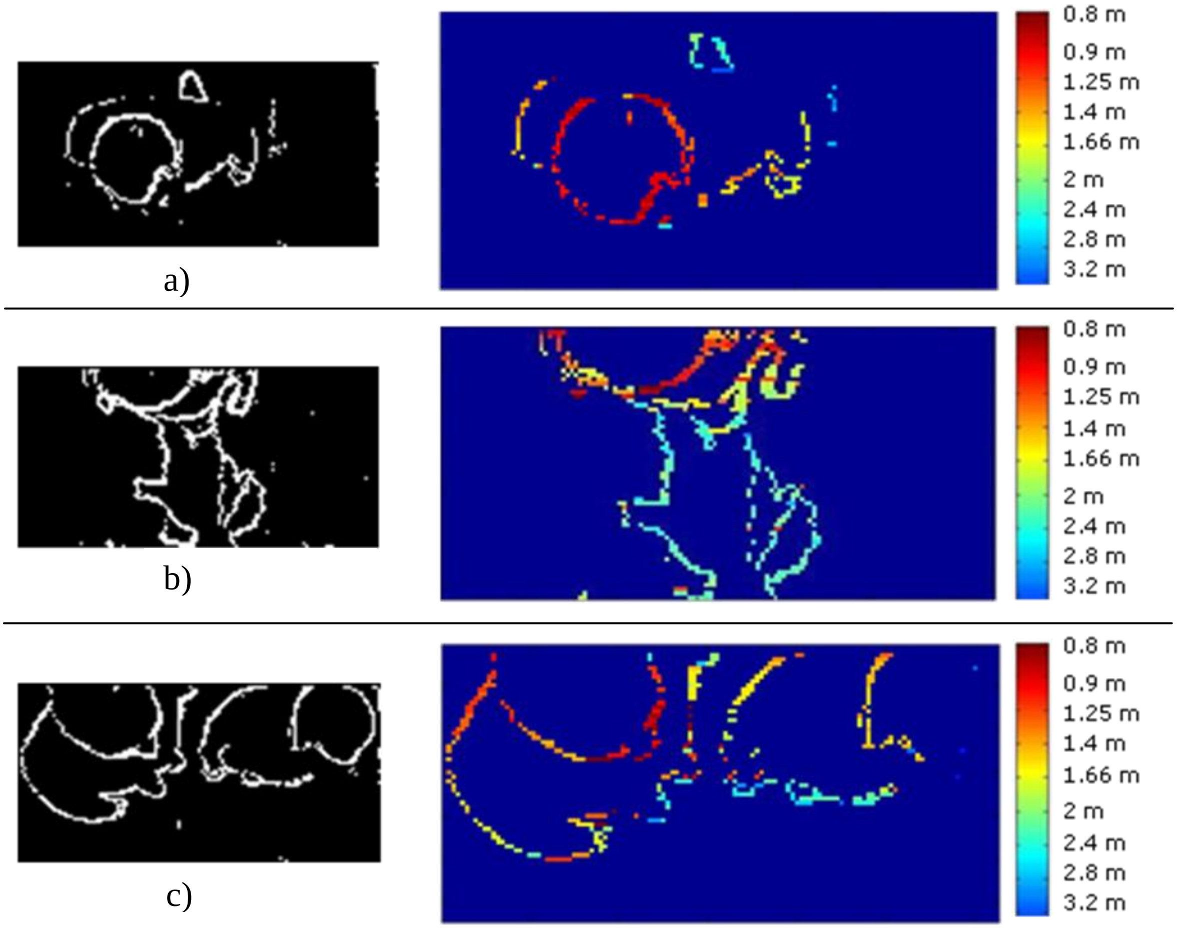


Figure 48. Prototype of the overhead people counting system. In the blow-up, the stereovision system is depicted with geometric details.

The system shows the image of scene and the disparity map respectively. Figure 49. Although we can see the edges are few, they are well defined and the corresponding results obtained are reasonably good. Starting from the different disparity value we can discriminate the different parts of the body such as shoulders, head and feet very clearly.



a)

b)

c)

Figure 49. Example of the disparity maps obtained. a) Identification of Heads, Shoulders and Feet b) Body and shadow discrimination c) Identification of different heads

The Figure 50. shows the different resolution relative to the distance from the sensor.

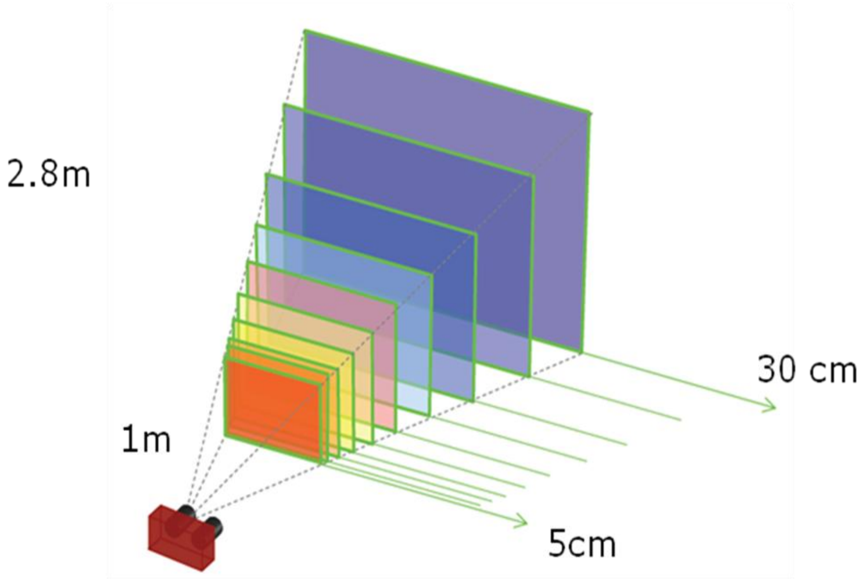


Figure 50. Depth resolution of stereo vision system realized.

These results can be considered as a feasibility study and demonstrate that it is the correct approach for this type o image. These results could be extended for use with other similar sensors and also with higher resolution sensors in order to obtain a better performance.

Although, these results have been obtained with a sensor having low pixel count, it can be ported to a similar architecture having much larger resolution. In this case, the depth precision is linearly related to the number of pixels mapped the scene.

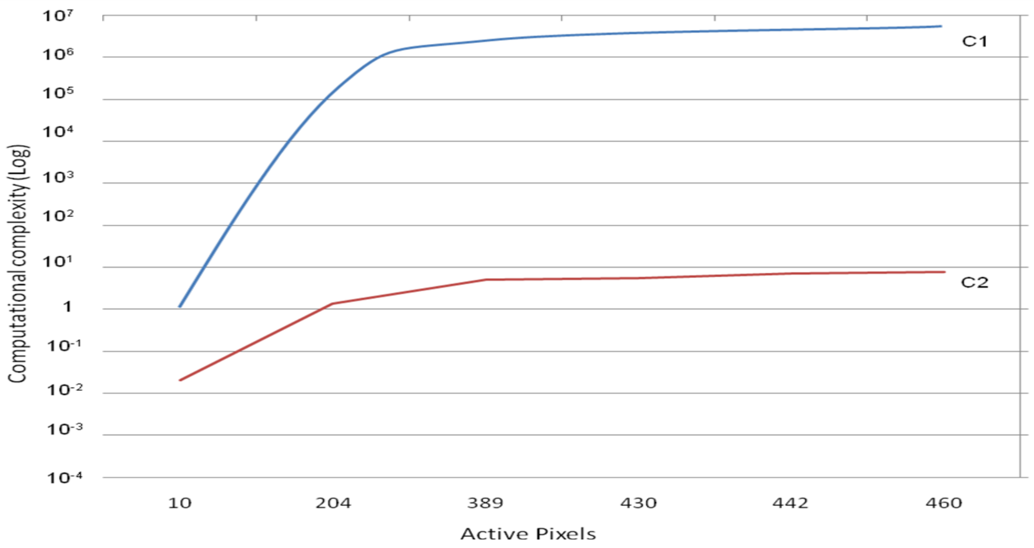


Figure 51. Analysis of algorithms and computational complexity

Figure 51. shows a graphic representation between the worst case (C1) complexity, with respect to the available active pixels in the image, and the real case (C2) referred to our benchmark. It is worth noting that C1 grows as C^4 with respect to the active pixels, while C2 is $O(C^2)$ to power two. This reduction is relate to the contents described in Figure 37.

In order to demonstrate the different computational complexity and the complex comparison of the custom sensor approach with respect to a standard approach, we refer to a common sensor resolution of 128 x 64 pixels. While a standard camera delivers grey-level data at 65Kbit/frame (8 bits/pixel), the spatial-contrast sensor delivers data at a maximum rate of 4Kbit/frame, turning into a down-scaling factor of 16 in the memory requirements with respect to the standard imager.

The proposed algorithm has the main advantage of dynamically adapt its computational load to the current amount of data delivered by the sensor. As specified above, the amount of data depends on the specific operating scenarios and can typically range from 0.1*N down to 0, where N is the total amount of pixels in the sensor.

The computational complexity necessary to execute the custom algorithm is O(C^2), where C is the number of active pixels, defined as C (contrast) =µ*N, where N is the sensor resolution and µ is the typical percentage of active pixels in the image (ranging between 0.05 and 0.1).

Table 4 compares the complexity of the algorithm proposed in this Thesis with that one based on full-block matching and a third one with modified matching, implemented on-chip by R.M. Philipp and R. Etienne-Cummings [54]. All of them refer to an imager with 128x64 pixels.

128*64 Pixels	Full block matching	Philipp and Cummings	This algorithm (10% active pixels)
Operations/ Frame	4917 Kops	450 Kops	83 Kops
Operation Reduction	1	1/11	1/60

Table 4. Number of operations comparison (ops) , general case full block Matching, Philipp and Cummings hardware implementation and this algorthm.

It is worth observing that the proposed algorithm drastically reduces the total amount of operations by a factor of about 60 times with respect to the full block matching, which represents here the worst case. Moreover, this approach, scales fairly well with the imager resolution and equally with respect to the power consumption.

Table 4 does not take into account another important feature of the spatial-contrast sensor. Assuming that a standard camera performs a dynamic range of about 60dB, additional processing needs to be done in order to reach the 100dB obtained with the spatial-contrast vision sensor. In this case, double-sampling technique can be adopted

[55] Double-sampling technique is based on a linear combination of two successive images acquired with two integration times. This means that, for each high-dynamic range frame, two images have to be acquired and read out, almost halving the frame rate and doubling the memory requirements and the power consumption. A first image (F1) is taken with a long exposure time T1. A second frame (F2) is acquired with a much shorter integration time (T2). The final image is the result of F = F1/k + F2; where k is a coefficient typically set to 2. By adopting this technique, the resulting dynamic range is given by T1/(2*T2).

4.5. Closing remarks

This part deals with stereo vision system a subclass of stereo vision in which the with the goal it’s to demonstrate the exploitation of custom CMOS sensor characterized by low power for stereo vision application.

We reassume this work in different steps, the first, to study and development of an efficient stereo-vision algorithm tailored on the spatial contrast information directly delivered by the CMOS sensors through a custom positional compression. The processing aims at extracting the disparity map with reduced computational resources and memory with respect to a standard approach. The algorithm has been simulated and evaluated on stereo datasets acquired with standard cameras before to be tested onto the custom CMOS sensor.

The second step is focused to Develop a prototype of stereo vision system based on the spatial contrast sensor. The two sensors are driven and interfaced with an FPGA linked to a PC through the USB. The stereo-vision algorithm runs on the PC and provides the disparity map at a frame rate of 15frames/s. The system has been preliminary tested on simple real scenarios. The whole system has shown a degree of accuracy in the calculation of the distances (an error of 10 cm over a distance of 2 m), which limited on the low resolution of the sensor.

Chapter 5

5. Ultra low power Vision system for scene interpretation

Taking advantage for the previous work where a contrast-based vision sensor has been analyzed and tested in real application scenarios, we came up with the considerations that poor binary contrast-based information is not enough for a reliable image processing. Although this information is properly provided over a wide intra-scene dynamic range, contrast is a high frequency spatial quantity which doesn't provide information inside of object. This is a very serious limit which make this class of sensors to be useful only in a limited and controlled class of application.

The idea is to study a new custom sensor able to implement the temporal contrast. With respect of spatial gradient the temporal contrast extract the entire silhouettes of moving object respect of the simple edge detections.

Other Big advantages for sensor integration due to the fact that the algorithm is based on pixel-level temporal processing. A low-level approach is considered, requiring no interactions with neighboring pixels. Pixels, which do not need interconnections with other pixels, are much simpler to design. In fact, interconnections are very expensive, they occupy large silicon area and does not scale very well with the technology.

The moving objects identification is one of fundamental and critical task in video surveillance or in human monitoring and analysis, detection and tracking, among other applications.

In general the background subtraction is a useful common approach to discriminate the moving objects in the scene, where each video frame is compared against a reference or static background model.

This second system developed is based on a dynamic frame difference implementation. In fact one of the problems of frame difference is inability to manage the background dynamically. Over time the standard frame difference extracts every new object present in the scene with respect to the background. This technique has a limitation because often the background can change over the time and the system is not able to define that a new periodic event should be absorbed in to the background.

Considering the human vision, if we observe the environment we focalize on new changes with respect to previous observations. If these changes are periodic the our brain absorb these changes to the background. The idea is to replicate this human capacity in a image sensor system.

The system development uses a custom sensor based on temporal difference principle seen before in chapter 2. It's equipped with a programmable filter able to perform a background suppression dynamically through the second level algorithm.

5.1. Custom chip embedded algorithm for scene interpretation

The idea is starting from an algorithm already tested an validated by a company EMZA that produces systems video surveillance and that it has patented this algorithm on project Bovis [56]. The benchmarking activity brought to the definition of an algorithm based on the Emza's algorithm and suitable for CMOS implementation.
This algorithm is based on two different threshold defined in this way:

$$Min_{i,j}(nT) = \min(Min_{i,j}(n-1)T, Pix_{i,j}(nT)) \quad (14)$$

$$Max_{i,j}(nT) = \min(Min_{i,j}(n-1)T, Pix_{i,j}(nT)) \quad (15)$$

$Pix_{i,j}$ (nT) is the current value *i,j*-th pixel at time *nT*, where *T* is the time between two image acquisitions and *n* is an integer number.
In (Figure 52.), an example of scene sequence, showing a person entering the scene. In this case, the floor has a regular pattern and the scene is partially illuminated by the sun. The illuminated zone slightly changes and moves along the image, causing a slowly changing scene. The algorithm adapts for this slowly changing illumination and can detect the moving person from background. After applying equations (14) and (15) and binarizing the resulting signal, the output of the algorithm is shown in (Figure 52.). Here, it possible to note that the background (dc signal) has been suppressed, while the slowly changing zone, generated by the sunlight (top side), has been almost completely rejected thanks to a sort of self-adapting capability implemented through a dynamic background subtraction.

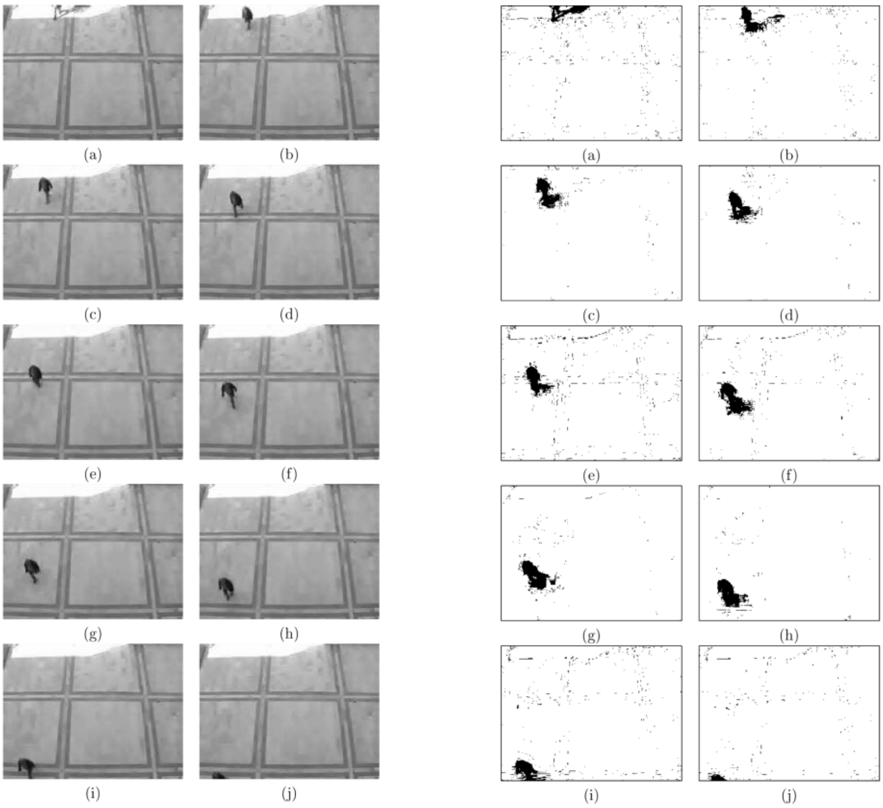


Figure 52. Real images and output of feature extraction algorithm

Figure 53	N. hot pixels
(a)	1793
(b)	1687
(c)	1502
(d)	1489
(e)	2310
(f)	2105
(g)	1783
(h)	1513
(i)	1901
(j)	704

Table 3 Number of hot-pixels present in the single images

In Table 3 the number of hot-pixels are reported for each image of (Figure 52.). Those images are taken with VGA resolution (640x 480) (i.e. 300Kpixels). In this case, the resulting binary images contain a very low percentage of hot-pixels, reaching no more than 0.7%. This turns into a highly compressed visual information, which needs to be post-processed. Moreover, if we consider that the source signal has 8-bits resolution, while the output signal has 2 bits/pixel, the latter ratio shrinks down to less than 0.2%.

After these preliminary verifications, the core algorithm has been slightly modified and adapted, according with the CMOS technology constraints.

The main idea behind this is that CMOS technology brings a significant improvement to the architecture only if the pixel parallel processing can be implemented at pixel-level with simple and compact electronic.

At a first glance, this can be done only through an analog design approach. A digital approach, as it was originally conceived and implemented by Emza, cannot be efficiently integrated into a CMOS sensor. Converting the photo-generated signal of the pixels outside the array, as most of the currently available CMOS imagers do, turns into a poor efficient solution in this case, maintaining the system complexity without exploiting the advantages of CMOS IC technology.

From an intuitive point of view, each pixel observes the scene and estimates the type of activity (min, max) along a proper observation time. In case no suspicious event has been detected, this type of activity has to be disregarded. The pixel works like a kind of band-pass filter, suppressing dc signals but also recurrent events. This is accomplished by changing the filter parameters according with the scene.

The algorithm, to be embedded in a CMOS sensor, requires three images (*Current Image (CurrI), Min Image (MinI), Max Image (MaxI)*) as inputs and generates one binary image (*BinI*) with 2 bits for each pixel. Working with 8-bit resolution, a N x M pixel imager will require 3 x N x M x 8 bits memory for the input, plus a 2 x N x M bits memory for the output.

The pixel requires 2 comparators and 2 binary-memory, storing Min (Max).

Due to the large amount of required memory, an analog implementation has been preferred to a digital one. One of the reasons is that analog can be better embedded into pixel together with comparators, comparing the current photo-generated signal with the past value. In this way, current and past values are both stored into analog memories, which can be implemented very close to each other, reducing parasitic effects, coupling effects and delay.

Now, we will describe some details of a pixel architecture which is intended to estimate the equations (14) and (15) and to binarize the result, according with the project specifications. The operating functions of this pixel have been previously simulated with MATLAB on real image sequences (Figure 52.) using Emza's low-level algorithm. After a preliminary verification of the results, the algorithm has been slightly modified and adapted in accordance with the constraints of a CMOS technology.

For example, usual mathematical operators, like product and division which are widely used in digital, have been avoided or implemented using alternative approaches or approximations, aimed at adopting simple operations between signals which can be easily implemented in CMOS: sum, subtraction, absolute value, thresholding, etc.

The activity and test of the low power vision system, have brought to the definition of a new pixel topology suitable for IC integration, which performs adaptive image pre-processing with reduced energy budget. This is intended to be the basic building block for a novel low-power vision sensor architecture with advanced performance.

It must be pointed out that although the low-level part of Emza algorithm requires massive pixel-parallel computations, it only uses simple operators like sums, subtractions and comparisons. These characteristics meet the constraints of the CMOS technology. In fact, only few simple operations can be efficiently integrated near the photodetector, maintaining a reduced pixel size.

5.2. Algorithm simulation

First the prototyping phase, the pixel electrical schematic has been designed and simulated. As we mentioned before, the algorithm will be implemented on chip generates a hot pixels only when the value of the current image exceeds a certain threshold. The objective of this algorithm is to manage the filter dynamically obtaining different behaviors depending on the application in which it is used.

The pixel embeds two analog memories (Max, Min) keeping trace of the photodiode activity along time. The two levels define a voltage range inside which the pixel activity is to be considered “normal”. Under a significant change in light intensity, the pixel voltage trespasses one of the two thresholds (VP > VMax “light → dark” or VP < VMin “dark → light”), setting itself into a suspicious state: “hot-pixel”. VMax and VMin are voltages with programmable time-constant, computed by two analog Switched-Capacitor Low-Pass Filters (SCLPF).

The SCLPF transfer function simulated is:

$$H(s) = \frac{1}{1+s\tau} = \frac{1}{1+s\left(\frac{C_{2M}}{C_{1M}}\cdot\frac{N}{f_o}\right)} \tag{16}$$

where C1M and C2M are the filter capacitors, with C2M/C1M = 1.3

After the exposure time, the current photodiode voltage (VP) is compared with both VMax and VMin, providing two-bits/pixel, which define one of the three allowed status (H [0,1]: VP > VMax ; M [0,0]: VMax > VP > VMin ; L [1,0]: VP < VMin).

This binary image is ready to be processed outside the chip by the higher-layer algorithm, implementing complex vision tasks with the different N values.

In order to understand the functioning of the system we analyze the Figure 53. Where a sequence of 20 frames is plotted.

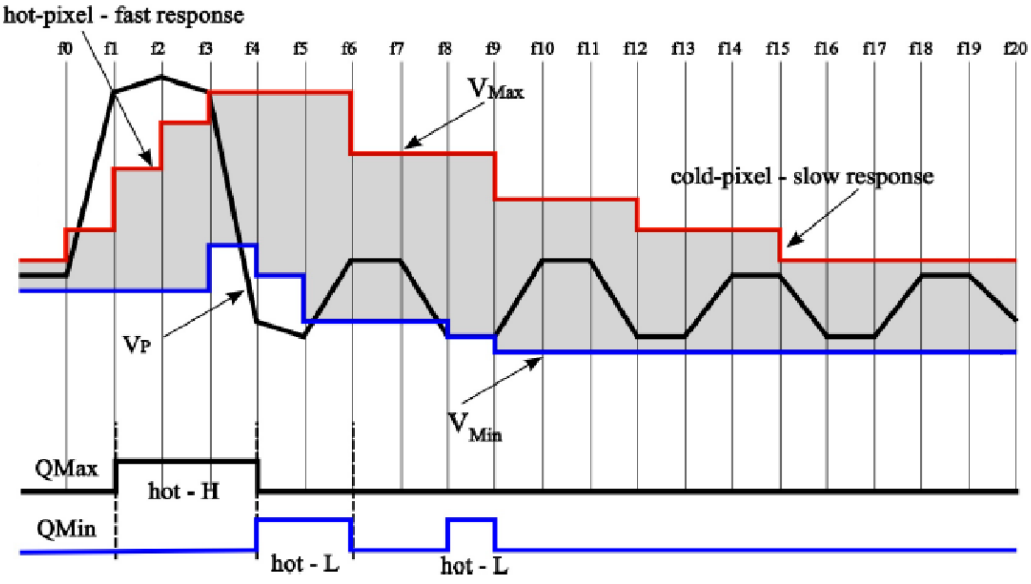


Figure 53. Funcitionality pixel algorthm

The algorithm implemented defines two signal thresholds V_{Max} and V_{Min} around the current value $V_{P.}$ The activity of the algorithm is to keep inside a boundary between V_{Max} and V_{Min} the V_p. These values change over time, adapting to the light intensity, if V_p is constant the V_{Max} and V_{Min} converge toward VP maintaining a minimum distance in order to absorb the V_p irrelevant variation and noise.

In the case $V_P > V_{Max}$ or $V_P < V_{Min}$ (hot pixel) the pixel is labeled as “hot pixel”, indicating a potential alert condition.

If V_P rapidly changes between two different values, the pixel decreases its sensitivity by separating V_{Max} and V_{Min}, as shown in part (Hot pixel) of Figure 53. The pixel state is encoded by two digital signals Q_{Max} and Q_{Min}, that are set high when the hot pixel is detected in a current situation and are updated at the end of each frame.

The objective of the algorithm is to maintain the thresholds around the current value VP where the pixel is considered normal “cold pixel”.

The sensor recognizes only the information of movement within the image. Using this technique it possible to absorb the movement dynamically and to maintain a low data delivery. When the hot pixel is present in the image of the objective of the pixel algorithm is to absorb a new event quickly. On the other hand when the event has already been identified no pixel is generated and the memory should be quickly updated.

As we can see in the Figure 53. starting from the data generated by the sensors it is possible to update the memory during the hot or cold pixel phase in an asymmetrical way through the N value manage by second level algorithm. This asymmetric updating of the memory allow use to manage the sensitivity or insensitivity of new events.

In fact in case the “hot-pixel” is not recognized to be associated with a suspicious event, its VMax or VMin are slowly updated toward the current value VP by means of the two SCLPFs, aiming at suppressing the “hot-pixel” status. Here, the pixel is desensitized with respect to next similar signal variations. If the “hot-pixel” is associated with a suspicious event, VMax and VMin are not updated. Therefore, for similar signals in next frames, the pixel will be still recognized as “hot-pixel”.

We selected On CAVIAR database the standard video scenarios characterized by frontal view of Shopping Center. Different movement condictions are presented in this database with a person going in to an out of a store, people walking together along the corridor and a person stopping outside a store,shown in Figure 54.

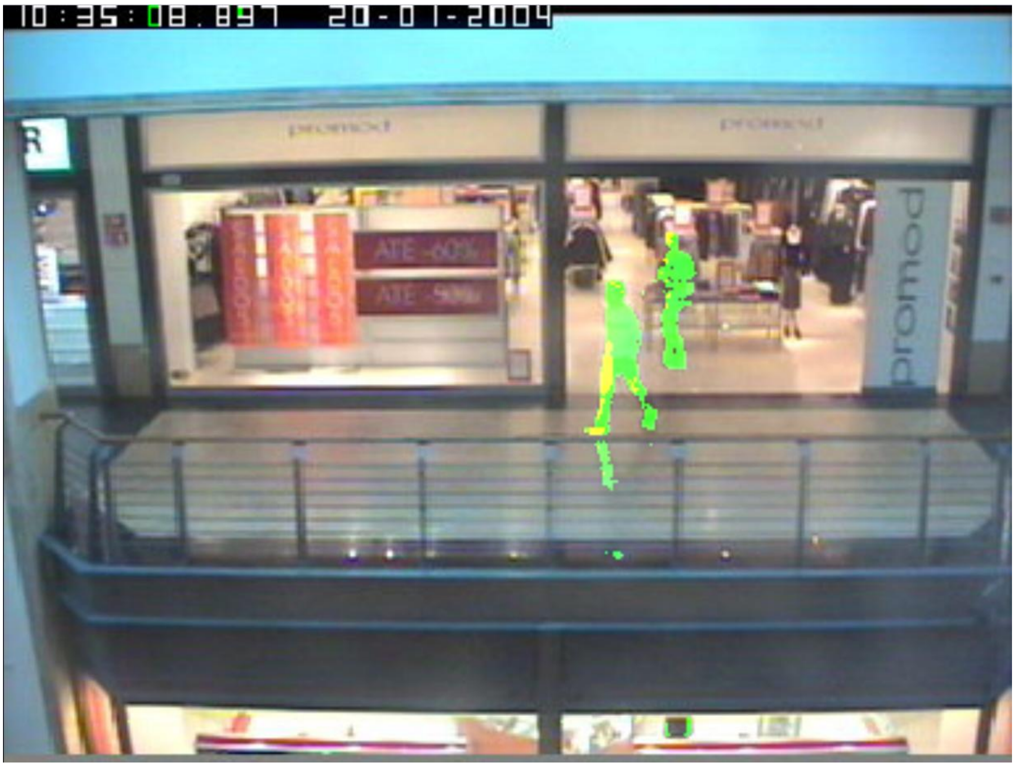


Figure 54. Output of pixel algorthm.Green rappresent the Q_{Max} and yellow the Q_{Min}

5.3. System realizations

The first measured that we have performed on the prototype fabricated is related on the transfer function characterization through the simple patter generator exploitation.
The test confer the SCLPF transfer function:

$$H(s) = \frac{1}{1+s\tau} = \frac{1}{1+s\left(\frac{C_{2M}}{C_{1M}} \cdot \frac{N}{f_o}\right)} \qquad (17)$$

where C1M and C2M are the filter capacitors, with C2M/C1M ~ 1.3, with some non ideality related to hardware implementation.

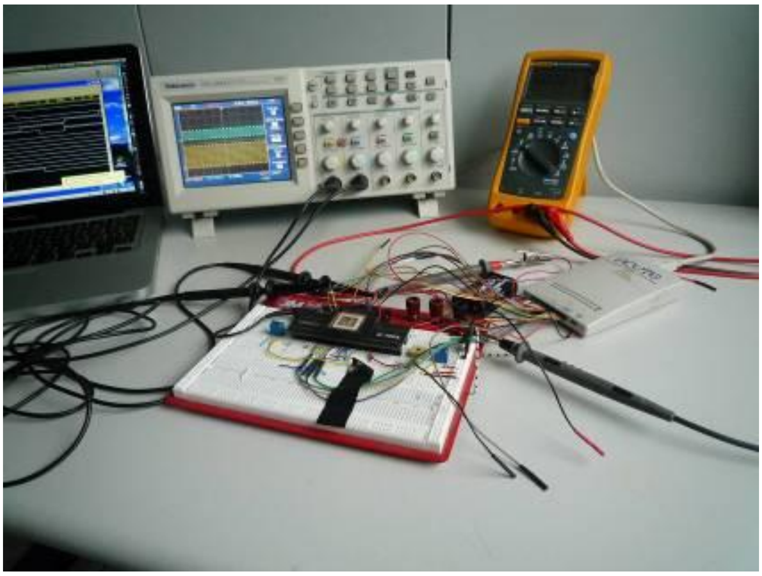
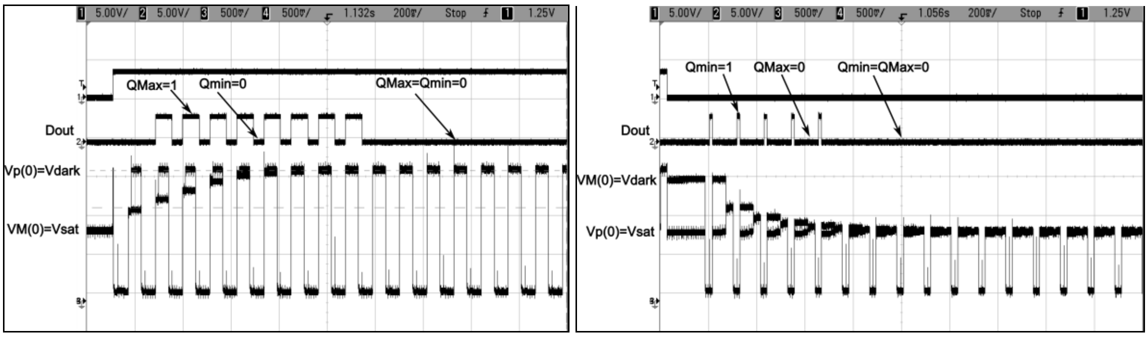


Figure 55. Prelinary funcitionality pixel test with pattern generator



a) b)

Figure 56. Evaluation τ parameter of the transfer function a) step-up b) step-down

Two basic pixel test operations are:

a) step-up, where the memory VM starts from Vsat and reaches the current value Vp=Vdark. The process requires 8 frames to settle from “hot-pixel” to “normal”;

b) step-down, where the memory Vm starts from the highest value (Vdark) and reaches the current value Vp=Vsat. The process takes 5 frames to settle.

Due to mismatch and capacitive coupling, the two processes do not have the same time constant. This is not really a problem, being the sensor in a feedback loop with a direct control on each pixel. After the hardware algorithm validation we made the prototype to test the sensor and the algorithm with real data. In order to do this the chip have been mounted on a custom PCB plugged onto a tiny FPGA development board. The FPGA provides the proper stimuli to the chip The FPGA is connected to a PC through a USB link. It reads out digital data from the chip and send them to a PC-based digital acquisition board. A Graphic User Interface (GUI) allows the user to change few sensor parameters (Integration time, SCLPF clock) and displays the binary “hot-pixels” images through a LabView.

FPGA generates the proper waveforms for the chip, and at the same time, acquires the three analog images (Vp, Vmax, VMin) together with the two binary images (2 bits/pixel). In this way, a complete monitoring of the chip functionalities can be accomplished. The present setup is very flexible and easily configurable, allowing to test the sensor under different driving configurations.

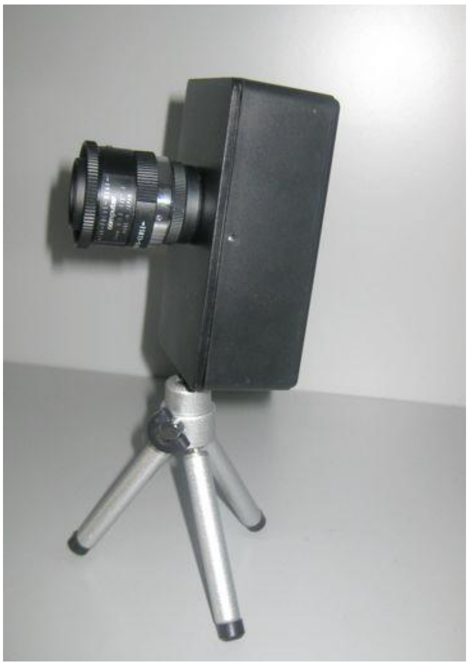


Figure 57. The prototype realized on FPGA platform usb link with a PC

In the debug phase in addition to the digital part it was also useful to extract the analog components of the memory to be able to understand clearly the behavior of the sensor.
In fact during the use of the this type of sensor it was interesting to observe and understand which where the instant values of the memory in order to determined the presence or not of the data.

5.4. System results

In order to make the functional test reproducible, a benchmark movie is projected onto a monitor and acquired by the vision sensor show in Figure 58. .

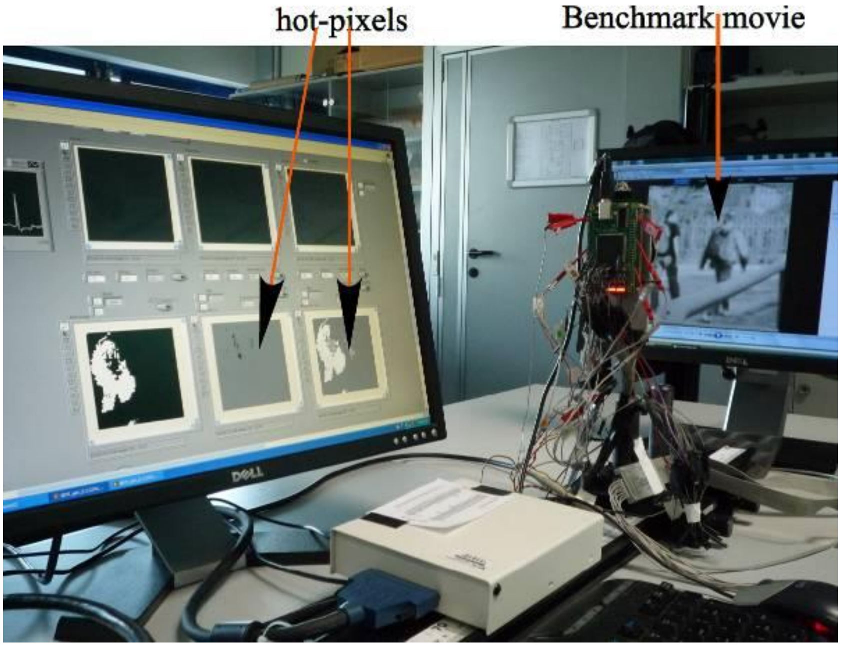


Figure 58. First measure of real benchmark movie.

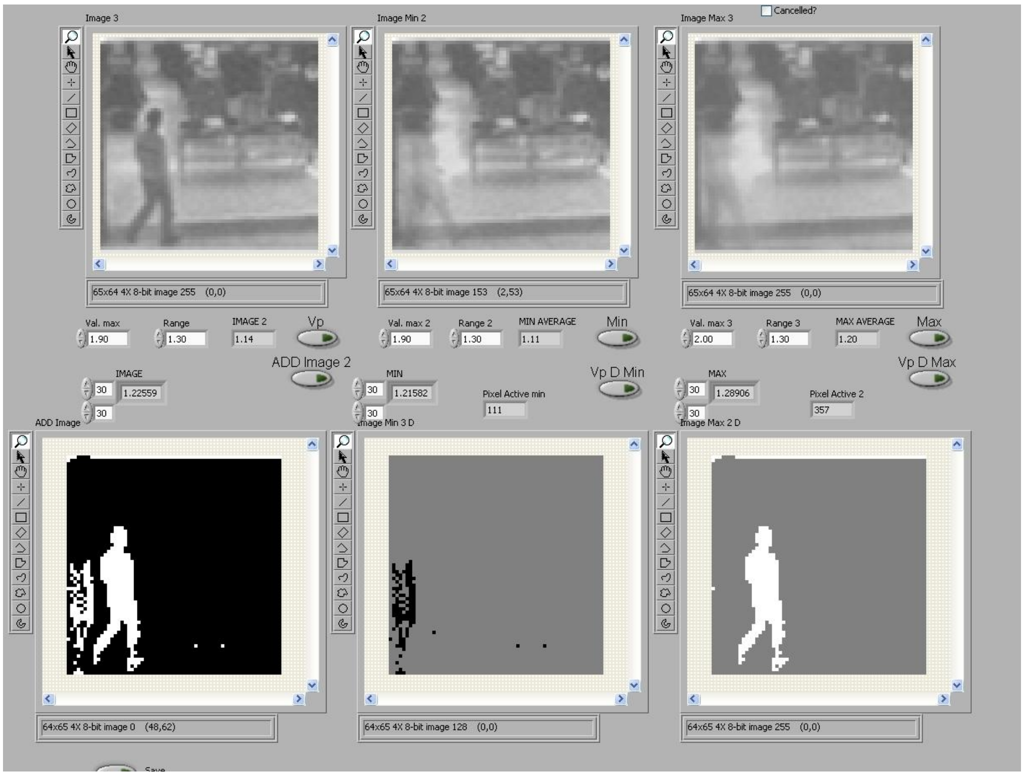
Different time of updating are used in order to understend the effects of the mouvments.

The chip extracts active pixels form the image related to moving patterns and provides tree binary images in Figure 59. :

Image Gray-Black — where the pixels measure changes from dark-to-light;

Image Gray-White — where the pixels measure changes from light-to dark;

Image Black-White — where the pixels measure changes;



65x64 4X 8-bit image 255 (0,0)
65x64 4X 8-bit image 153 (2,53)
65x64 4X 8-bit image 255 (0,0)

Val. max 1.90
Range 1.30
IMAGE 2 1.14
Val. max 2 1.90
Range 2 1.30
MIN AVERAGE 1.11
Val. max 3 2.00
Range 3 1.30
MAX AVERAGE 1.20

IMAGE 30 30 1.22559
MIN 30 30 1.21582
Pixel Active min 111
MAX 30 30 1.28906
Pixel Active 2 357

64x65 4X 8-bit image 0 (48,62)
64x65 4X 8-bit image 128 (0,0)
64x65 4X 8-bit image 255 (0,0)

Figure 59. Labview interface. On top Vp,Vmin e Vmax abalog values are ripectivly rapprest. At the bottom Digital imgage changes,min memory changes, max memory changes

The system reveals only movement relative to the people and not to the background.

The other three picture on top represent the analog value, useful only for debug, related to, current image, minimum memory and the maximum memory

The vision sensor has been tested forcing a different time response of the SCLPFs, in order to verify the proper operating modes. In Figure 60. shows the response of the SCLPF1 and SCLPF2, during this test the sensors is rapidly exposed to a high light an when the memories are stable to a dark light.

Under an “hot-pixel” the filter is clocked once every frame, forcing the SCLPF to have the fastest response, in order to rapidly compensate for anomalous situations. Under a “normal pixel” the SCLPF is clocked once every two frames, slowing down the filter by a factor 2, producing a persistence effect in the vision sensor response.

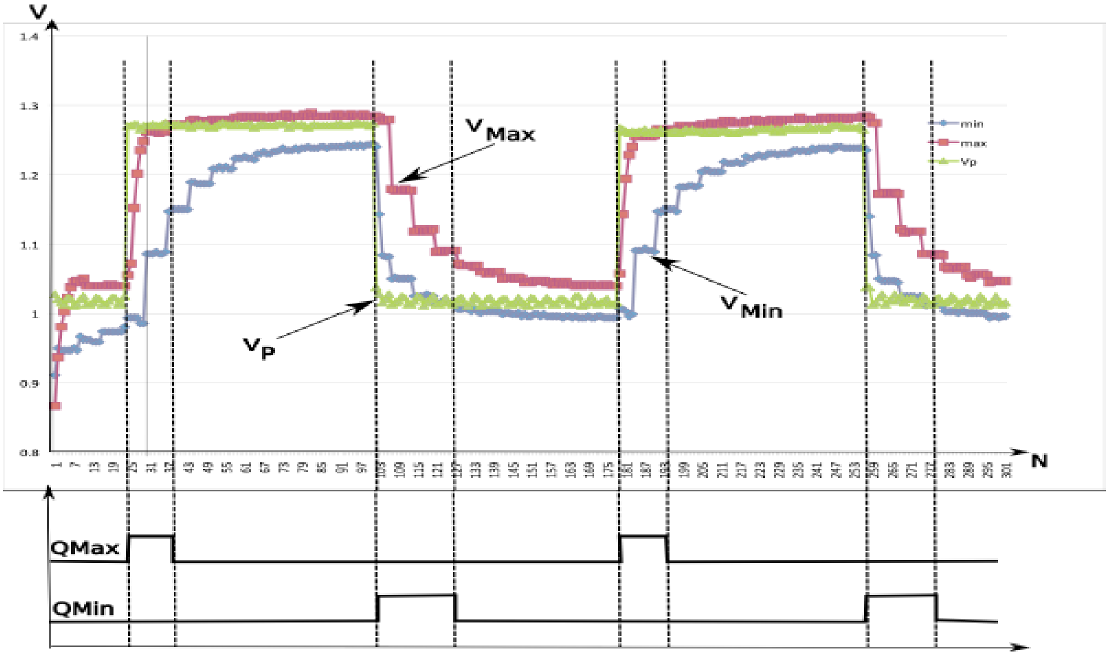


Figure 60. Temporal respond of the SCLPF1 and SCLPF2 to light changes.

Through different updating of the filter it is possible to parameterize the speed of the absorption of the movement. In other word this turns into programmable local area filter of a second level algorithm.

5.5. Closing remarks

The main benefit of visual processing, based on spatial contrast, is the low complexity, thanks to the reduced amount of data involved in the processing.

On the other hand, its most critical drawback is given by the strong relationship between the sensitivity of spatial gradient estimation and the size of the pixel kernel, which is usually wired on an ASIC implementation. A programmable kernel would be therefore necessary but practically not efficiently implementable on-chip.

A much better approach, which is also efficient and easy to be embedded on silicon, is the temporal contrast estimation. In fact, it does not require any kernel-based operations, which are very expensive to be implemented in hardware. It only needs one or few memories per pixel to store previous values to be related to the current one. The basic operations involved in this approach are storage, sums and comparisons; very easy to be implemented in CMOS.

Based on this computational paradigm, a working vision sensor prototype has been built, embedding dynamic background subtraction. The sensor is able to detect anomalous events in the scene by removing all static pixels and also those pixels with slow or regular intensity changes.

6. Conclusions

This PhD dissertation is focused on the study and the development new low power vision systems architectures, which aim at closing their gap with the energy autonomous systems. Stating that custom hardware is the best approach to follow in this application area, my work was organized in order to investigate novel computational paradigms for visual processing aimed at exploiting the potentials of two vision sensors, targeted to ultra-low power applications.

A custom stereo-vision algorithm has been developed and evaluated, tailored to a contrast vision sensor, delivering binary data, which are compressed by means of a positional coding. In contrast to a standard stereo processing, based on intensive pattern matching, the new algorithm exploits the sensor data coding with no redundancy, taking advantage to dynamic programming. Good experimental results have been obtained although the system has been tested in simple application scenarios. For example, a 5cm depth resolution has been obtained, it demonstrate the correct approach adopted, for this type of senor.

The most important critical points depend on the spatial contrast extraction method, which is embedded in the sensor. This method is implemented in an hardwired kernel, i.e. it is tuned on a specific range of spatial contrast values, with no possibilities to be changed. This makes the sensor to have different sensitivities with respect to different scenarios. Spatial kernel programmability, would be extremely expensive if implemented at sensor-level. It would turn into a large pixel size with complex connectivity among.

Much better performance could be obtained by using adaptive temporal contrast rather than spatial contrast information. In fact, temporal contrast does not require spatial connectivity and can be efficiently embedded into silicon together with simple adaptive processing to make each pixel to change its own sensitivity upon request.

The adaptive temporal-based approach has been firstly simulated and a novel architecture of vision sensor has been defined, which partially embeds pixel-level parallel processing.

A new vision sensor has been developed and fully tested. The sensor directly extracts a binary image where the active pixels are those detecting an anomalous signal variation with respect to the previous behavior. If compared with respect to the first sensor, here the binary information is only related to the motion, as expected, and its sensitivity to is much larger, Each pixel work around its maximum allowable sensitivity, which makes it to be very sensitive to any signal variations, in accordance with an adaptive principle. A system demonstrator based on the temporal contrast sensor has been used and validated in different indoor scenarios.

Future work and perspectives are related to combine different computational distribution of standard acquired system. Although the two sensors have some common characteristics, like binary output data, they are based on different principles for features extraction. This means that new algorithms and processing approaches need to be investigated and developed for the new sensor architecture emphasizing the adaptive peculiarity of the system. In other words, the sensor needs to be put into a feedback loop with the processing unit.

Temporal contrast demonstrated to be a very reliable approach which is also suitable to be efficiently integrated at sensor level. This would pave the way of a new class of energy-aware vision sensors based on events detection and targeted to monitoring applications. Much work needs therefore to be done at hardware-level, investigating novel CMOS sensor architectures taking advantages from new hardware-oriented low-level algorithms.

Bibliography

[1] J. Marienborg, T. Lande, A. Abusland, and M. Hovin, “An analog approach to ”neuromorphic” communication,” IEEE ISCAS 96, vol. 3, pp. 397–400, 1996, atlanta, GA

[2] Yuri A. Ivanov and Aaron F. Bobick, “Recognition of Visual Activities and Interactions by Stochastic Parsing”, IEEE Trans. Pattern Anal. Mach. Intell., Volume 22, Number 8, 2000, pages 852-872.

[3] [Nemzek04] R. J. Nemzek and J. Dreicer, “Distributed sensor networks for detection of mobile radioactive sources,” IEEE Transactions on Nuclear Science, vol. 51, pp. 1693–1701, August 2004.

[4] K. Morioka, J. Lee, and H. Hashimoto, “Human-following mobile robot in a distributed intelligent sensor network,” IEEE Transactions on Industrial Electronics, vol. 51, pp. 229–238, February 2004.

[5] E. Magli, C.F. Chiasserini, “Energy-efficient coding and error control for wireless video surveillance networks,” in Telecommunication Systems, vol. 26, pp. 369–387, JUN-AUG 2004.

[6] Y. Liu, W. Gao, H. Yao, S. Liu, and L. Wang, “Fast moving region detection scheme in ad hoc sensor network,” in Proceedings Lecture Notes In Computer Science, vol. 3212, pp. 520–527, 2004.

[7] Weiming Hu and Tieniu Tan and Liang Wang and S. Maybank, “A survey on visual surveillance of object motion and behaviors”, IEEE Transactions on Systems, Man and Cybernetics Part C, volume 34, number 3, August 2004, pages 334-352.

[8] M. Duarte and H. Yu, “Vehicle classification in distributed sensor networks,” Journal of Parallel and Distributed Computing, vol. 64, pp. 826–839, July 2004.

[9] W.-C. Feng, B. Code, E. Kaiser, M. Shea, and L. Bavoil, “Panoptes: scalable low-power video sensor networking technologies,” Proceedings of the ACM International Multimedia Conference, pp. 562-571, Berkeley (CA), Nov. 2003.

[10] http://wsnl.stanford.edu/smartcam.php.

[11] C. B. Margi, X. Lu, G. Zhang, G. Stanek, R. Manduchi, and K. Obraczka, ”Meerkats: A power-aware, selfmanaging wireless camera network for wide area monitoring,” Workshop on Distributed Smart Cameras (DSC’06), October 2006.

[12] M. Rahimi, R. Baer, O. I. Iroezi, J.C. Garcia, J. Warrior, D. Estrin, M. Srivastava, " Cyclops: In Situ Image Sensong and Interpretation in Wireless Sensor Networks," Proceedings of the 3rd international conference on Embedded networked sensor systems, pp. 192-204, San Diego 2005.

[13] L. Ferrigno, S. Marano, V. Paciello, and A. Pietrosanto, “Balancing computational and transmission power consumption in wireless image sensor networks,” IEEE Int. Conf. on Virtual Environment, Human-Computer Interfaces an measurements, 2005.

[14] T. Teixeira, E. Culurciello, J.H.Park, D. Lymberopoulos, A. Barton-Sweney, and A. Savvides, “Address-Event Imagers for Sensor Networks: Avaluation and Modeling,” Proc. IPSN 2006, pp. 458-466, April 2006.

[15] http:// www.anafocus.com

[16] Ángel Rodríguez-Vázquez, Rafael Domínguez-Castro, Francisco Jiménez-Garrido, Sergio Morillas, Juan Listán, Luis Alba, Cayetana Utrera, Servando

Espejo and Rafael Romay "The Eye-RIS CMOS Vision System" in *Analog Circuit Design,* Springer Netherlands, 2008.

[17] K. Kagawa, S. Shishido, M. Nunoshita, and J. Ohta, "A 3.6 pW/frame·pixel 1.35 V PWM CMOS imager with dynamic pixel readout and no static bias current," in *IEEE Solid-State Circuits Conf. Dig. Tech. Papers*, 2008, pp. 54–55.

[18] M. Gottardi, N. Massari, and S. Arsalan Jawed, "A 100μW 128 x64 pixels contrast-based asynchronous binary vision sensor for sensor networks applications," *IEEE J. Solid-State Circuits*, vol. 44, no. 5, pp. 1582–1592, May 2009.

[19] Z. Fu and E. Culurciello, "A 1.2 mw CMOS temporal-difference image sensor for sensor networks," in Proc. IEEE Int. Symp. on Circuits Syst. *(ISCAS 2008),* pp. 1064–1067.

[20] S. Hanson and D. Sylvester, "A 0.45–0.7 Vsub-microwatt CMOS image sensor for ultra-low power applications," in *2009 Symp VLSI Circuits*, 2009, pp. 176–177, IEEE.

[21] K. Cho, D. Lee, J. Lee, and G. Han, "Sub-1-V CMOS image sensor using time-based readout circuit," *IEEE Trans. Electron Devices,* vol 57, no. 1, pp. 222–227, 2010.

[22] F. Tang, Y. Cao, and A. Bermak, "An ultra-low power current-mode CMOS image sensor with energy harvesting capability," in *2010 Proc. ESSCIRC*, 2010, pp. 126–129, IEEE.

[23] M. Law, A. Bermak, and C. Shi, "A low-power energy-harvesting logarithmic CMOS image sensor with reconfigurable resolution using two-level

quantization scheme,” *IEEE Trans. Circuits Syst. II, Express Briefs*, no. 99, pp. 1–5, 2011.

[24] R. Fontaine. Chipworks, Canada “A Review of the 1.4 mm Pixel Generation” Intl. Image Sensor Workshop 2011 IISW

[25] T. Teixeira, D. Lymberopoulos, E. Culurciello, Y. Aloimonos, and A. Savvides. A lightweight camera sensor network operating on symbolic information. In ACM SenSys Workshop on Distributed Smart Cameras. Citeseer, 2006.

[26] C. Mead, and M. Mahowald, A Silicon Model of Early Visual Processing. Pergamon Press, 1988.

[27] M. Sivillotti, ”Wiring considerations in analog VSI systems with applications to field programmable networks.” Ph.D. dissertation, California Institute of Technology, 1991.

[28] K. A. Boahen, ”Point-to-point connectivity between neuromorphic chips using address events,” IEEE Trans. Circuits and Systems-II: Analog and Digital Signal Processing, vol.47, no.5,pp. 416-434, 2000.

[29] E. Culurciello, R. Etienne-Cummings, and K. A. Boahen, ”A biomorphic digital image sensor,” *Journal of Solid- State Circuits*, pp. 281-294, Vol. 38, n. 2, 2003.

[30] E. Culurciello, R. Etienne-Cummings, “Second generation of high dynamic range, arbitrated digital imager, ”in IEEE Int. Symposium on Circuits and Systems, ISCAS, vol.4, Vancouver, Canada, 2004, 2004, pp. IV-828-31.

[31] A. Andreou, and K. Boahen, ”A 590,000 transistor, 48,000 pixel contrast sensitive, edge enhancing CMOS imager-silicon retina,” in Proc. Of the 16th

Conference on Advanced Research in VLSI, Chapel Hill, NC, 1995, pp. 225-240.

[32] E. Culurciello, and A. Andreou, ”CMOS image sensor for sensor networks,” Journal of Analog Integrated Circuits and Signal Processing, pp. 39-51, 2006.

[33] U. Mallik, M. Clapp, E. Choi, G. Cauwenberghs, and R. Etienne-Cummings, “Temporal Chance Threshold Detection Imager,” *ISSCC Dig. Tech. Papers*, pp. 362-363, Feb. 2005.

[34] Y. M. Chi, R. Etienne-Cummings, G. Cauwenberghs, P. Carpenter, K. Colling, “Video Sensor Node for Low- Power Ad-hoc Wireless Networks,” *Proc. ISCAS*, pp. 244-247, 2007.

[35] E. Culurciello, R. Etienne-Cummings, and K. A. Boahen, ”A biomorphic digital image sensor,” *Journal of Solid- State Circuits*, pp. 281-294, Vol. 38, n. 2, 2003.

[36] G. B. Zhang, T.H. Yang, S. Gregori, J. Liu, and F. Maloberti, “Ultra-low Power Motion-triggered Image Sensor for Distributed Wireless Sensor Network,” in Proc. IEEE Sensors Conference, Oct. 2003, pp. 1141 – 1146.

[37] P. Lichtsteiner, C. Posch and T. Delbruck, *A 128×128 120 dB 30 mW asynchronous vision sensor* that *responds to relative intensity change*, ISSCC (2006)

[38] Gruev, V.; Etienne-Cummings, R., "A pipelined temporal difference imager," *IEEE Journal of Solid-State Circuits* 39, 538-543 (2004).

[39] Dongsoo Kim; Zhengming Fu; Joon Hyuk Park; Culurciello, E.; , "A 1-mW CMOS Temporal-Difference AER Sensor for Wireless Sensor Networks,"

Electron Devices, IEEE Transactions on , vol.56, no.11, pp.2586-2593, Nov. 2009

[40] P. Lichtsteiner and T. Delbruck “64x64 Event-Driven Logarithmic Temporal Derivative Silicon Retina,” 2005 IEEE Workshop on Charge Coupled Devices and Advanced Image Sensors, June 9-11, Nagano, Japan.

[41] P.-F. Ruedi, “A 128 _128 pixel 120 dB dynamic range vision sensor chip for image contrast and orientation extraction,” *IEEE J. Solid-State Circuits*, vol. 38, no. 12, pp. 2325–2333, Dec. 2003.

[42] Dongsoo Kim and E. Culurciello, “A Compact-pixel Tri-mode Vision Sensor,” International Symposium on Circuits and Systems 2010, May 2010.

[43] Massari, N.; Gottardi, M.; , "A 100 dB Dynamic-Range CMOS Vision Sensor With Programmable Image Processing and Global Feature Extraction," *Solid-State Circuits, IEEE Journal of* , vol.42, no.3, pp.647-657, March 2007

[44] D. Scharstein, R. Szeliski, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,” International J. of Computer Vision, Vol 47(1-3), pp. 7- 42, 2002.

[45] Belbachir, A.N.; Schraml, S.; Nowakowska, A.; , "Event-driven stereo vision for fall detection," *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on* , vol., no., pp.78-83, 20-25 June 2011

[46] S. Schraml, A.N. Belbachir, N. Milosevic and P. Schoen,“Dynamic Stereo Vision for Real-time Tracking,” in Proc. of IEEE ISCAS, June 2010.

[47] Graph Theory, Book, by Reinhard Diestel.

[48] Dijkstra, E.W. (1959). A note on two problems in connection with graphs. Numer. Math., 1, 269-271.

[49] J.A.Bondy and U. S R. Murty, Graph Theory with applications, Elsevier North-Holland, 1976.

[50] An Ultra-Low-Power Contrast-Based Integrated Camera Node and its Application as a People Counter

[51] www.opalkelly.com/library/XEM3001-PB.pdf

[52] E. Trucco, A. Verri "Introductory Techniques for 3-D Computer Vision", Prentice Hall, 1998.

[53] "Camera Calibration Toolbox", www.vision.caltech.edu/bouguetj/calib_doc/

[54] R. M. Philipp and R. Etienne-Cummings, "Single-chip stereo imager," Analog Integrat. Circuits Signal Process., vol. 39, pp. 237–250, Jun. 2004.

[55] Y. Muramatsu, S. Kurosawa, M. Furumiya, H. Ohkubo, and Y. Nakashiba, "A Signal-Processing CMOS Image Sensor Using a Simple Analog Operation", IEEE J. of Solid-State Circuits, vol. 38, no. 1, pp. 101-106, January 2003.

[56] http://bovis.fbk.eu/node/1

[57] http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/

[58] Edsger W. Dijkstra. A note on two problem in connexion with graphs. Numerische Mathematik, 1:269–271, 1959.