Hindawi Publishing Corporation
EURASIP Journal on Embedded Systems
Volume 2009, Article ID 716317, 15 pages
doi:10.1155/2009/716317

Research Article

A System for an Accurate 3D Reconstruction in Video Endoscopy Capsule

Anthony Kolar,[1] Olivier Romain,[1] Jade Ayoub,[1] David Faura,[1] Sylvain Viateur,[1] Bertrand Granado,[2] and Tarik Graba[3]

[1] *Département SOC—LIP6, Universite P&M CURIE—Paris VI, Equipe SYEL, 4 place Jussieu, 75252 Paris, France*
[2] *ETIS, CNRS/ENSEA/Université de Cergy-Pontoise, 95000 Cergy, France*
[3] *Electronique des systèmes numériques complexe, Telecom ParisTech, 46 rue Barrault, 75252 Paris, France*

Correspondence should be addressed to Anthony Kolar, anthony.kolar@free.fr

Received 15 March 2009; Revised 9 July 2009; Accepted 12 October 2009

Recommended by Ahmet T. Erdogan

Since few years, the gastroenterologic examinations could have been realised by wireless video capsules. Although the images make it possible to analyse some diseases, the diagnosis could be improved by the use of the 3D Imaging techniques implemented in the video capsule. The work presented here is related to Cyclope, an embedded active vision system that is able to give in real time both 3D information and texture. The challenge is to realise this integrated sensor with constraints on size, consumption, and computational resources with inherent limitation of video capsule. In this paper, we present the hardware and software development of a wireless multispectral vision sensor which allows to transmit, a 3D reconstruction of a scene in realtime. multispectral acquisitions grab both texture and IR pattern images at least at 25 frames/s separately. The different Intellectual Properties designed allow to compute specifics algorithms in real time while keeping accuracy computation. We present experimental results with the realization of a large-scale demonstrator using an SOPC prototyping board.

1. Introduction

Examination of the whole gastrointestinal tract represents a challenge for endoscopists due to its length and inaccessibility using natural orifices. Moreover, radiologic techniques are relatively insensitive for diminutive, flat, infiltrative, or inflammatory lesions of the small bowel. Since 1994, video capsules (VCEs) [1, 2] have been developed to allow direct examination of this inaccessible part of the gastrointestinal tract and to help doctors to find the cause of symptoms such as stomach pain, disease of Crohn, diarrhoea, weight loss, rectal bleeding, and anaemia.

The Pillcam video capsule designed by Given Imaging Company is the most popular of them. This autonomous embedded system allows acquiring about 50 000 images of gastrointestinal tract during more than twelve hours of an analysis. The off-line image processing and its interpretation by the practitioner permit to determine the origin of the disease. However, recent benchmark [3] published shows some limitations on this video capsule as the quality of images and the inaccuracy on the size of the polyps. Accuracy is a real need because the practitioner makes an ablation of a polyp only if it exceeds a minimum size. Actually the polyp size is estimated by practitioner's experience with more or less error for one practitioner to another. One of the solutions could be to use techniques of 3D imagery, either directly in the video capsule or on a remote computer.

This later solution is actually used in the Pillcam capsule by using the 2–4 images that are taken per second and stored wirelessly in a recorder that is worn around the waist. 3D processing is performed off-line from the estimation of the displacement of the capsule. However, the speed of video-capsule is not constant; for example, in the oesophagus, it is of 1.44 m/s, and in the stomach it is almost null and is 0.6 m/s in the intestine. Consequently, by taking images at frequencies constant, certain areas of the transit will not be rebuilt. Moreover, the regular transmission of the images by the body consumes too much energy and limits the autonomy of the video capsules to 10 hours. Ideally, the quantity of information to be transmitted must be reduced

at the only pertinent information like polyps or other 3D objects. The first development necessary to the delivery of such objects relies on the use of algorithm of pattern recognition on 3D information inside the video capsule.

The introduction of 3D reconstruction techniques inside a video capsule needs to define a new system that takes into account the hard constraints of size, low power consumption, and processing time. The most common 3D reconstruction techniques are those based on passive or active stereoscopic vision methods, where image sensors are used to provide the necessary information to retrieve the depth. Passive method consists of taking at least two images of a scene at two different points of view. Unfortunately using this method, only particular points, with high gradient or high texture, can be detected [4]. The active stereo-vision methods offer an alternative approach when processing time is critical. They consist in replacing one of the two cameras by a projection system which delivers a pattern composed by a set of structured rays. In this latter case, only an image of the deformation of the pattern by the scene is necessary to reconstruct a 3D image. Many implementations based on active stereo-vision have been realised in the past [5, 6] and provided significant results on desktop computers. Generally, these implementations have been developed to reconstruct 3D large objects as building [7–14].

In our research work, we have focused on an integrated 3D active vision sensor: "Cyclope." The concept of this sensor was first described in [4]. In this new article we focus on the presentation of our first prototype which includes the instrumentation and processing blocks. This sensor allows making in real time a 3D reconstruction taking into account the size and power consumption constraints of embedded systems [15]. It can be used in wireless video capsules or wireless sensor networks. In the case of video capsule in order to be comfortable for the patient, the results could be stored in a recorder around the waist. It is based on a multispectral acquisition that must facilitate the delivery of a 3D textured reconstruction in real time (25 images by second).

This paper is organised as follows, Section 2 describes briefly Cyclope and deals with the principles of the active stereo-vision system and 3D reconstruction method. In Section 3 we present our original multispectral acquisition. In Section 4 we present the implementation of the optical correction developed to correct the lens distortion. Section 5 deals with the implementation of a new thresholding and labelling methods. In Sections 6 and 7, we present the processing of matching in order to give a 3D representation of the scene. Section 8 deals with wireless communication consideration. Finally, before a conclusion and perspectives of this work, we present, in Section 9, a first functional prototype and its performances which attest the feasibility of this original approach.

2. Cyclope

2.1. Overview of the Architecture. Cyclope is an integrated wireless 3D vision system based on active stereo-vision technique. It uses many different algorithms to increase accuracy and reduce processing time. For this purpose, the sensor is composed of three blocks (see Figure 1).

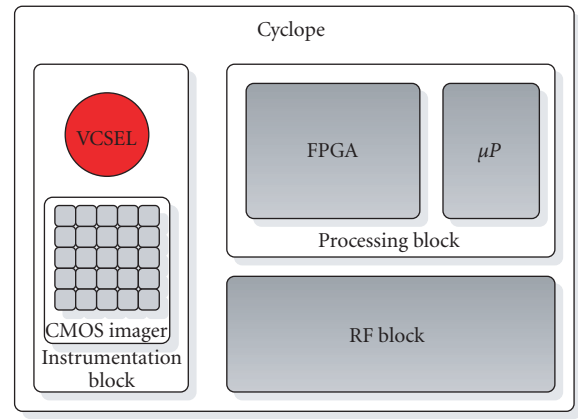


Figure 1: Cyclope Diagram.

(i) Instrumentation block: it is composed of a CMOS camera and a structured light projector on IR band.

(ii) Processing block: it integrates a microprocessor core and a reconfigurable array. The microprocessor is used for sequential processing. The reconfigurable array is used to implement parallels algorithms.

(iii) RF block: it is dedicated for the OTA (Over the Air) communications.

The feasibility of Cyclope was studied by an implementation on an SOPC (System On Programmable Chip) target. These three parts will be realised in different technologies: CMOS for the image sensor and the processing units, GaAs for the pattern projector, and RF–CMOS for the communication unit. The development of such integrated "SIP" (System In Package) is actually the best solution to overcome the technological constraints and realise a chip scale package. This solution is used in several embedded sensors such as The "Human++" platform [16] or Smart Dust [17].

2.2. Principle of the 3D Reconstruction. The basic principle of 3D reconstruction is the triangulation. Knowing the distance between two cameras (or the various positions of the same camera) and defining of the line of views, one passing by the center of camera and the other by the object, we can find the object distance.

The active 3D reconstruction is a method aiming to increase the accuracy of the 3D reconstruction by the projection on the scene of a structured pattern. The matching is largely simplified because the points of interest in the image needed to the reconstruction are obtained by the extraction of the pattern; it also has the effect to increase the speed of processing.

The setup of active stereo-vision system is represented in Figure 2. The distance between the camera and the laser projector is fixed. The projection of the laser beams on a plane gives an IR spots matrix.

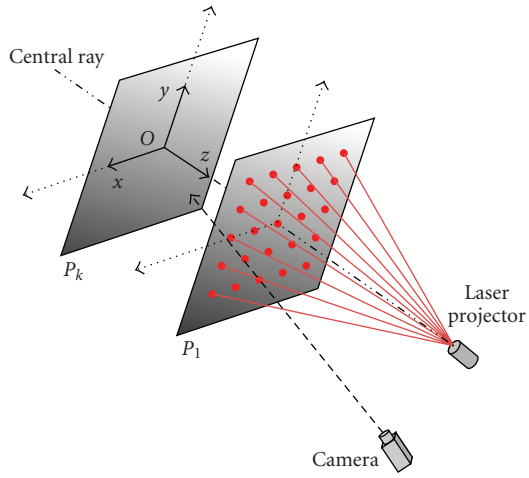


Figure 2: Active stereo-vision system.

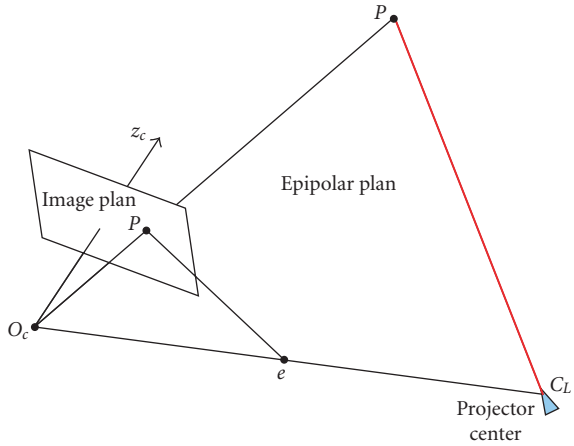


Figure 3: Epipolar projection.

The 3D reconstruction is achieved through triangulation between laser and camera. Each point of the projected pattern on the scene represents the intersection of two lines (Figure 3):

(i) the line of sight, passing through the pattern point on the scene and its projection in the image plan,

(ii) the laser ray, starting from the projection center and passing through the chosen pattern point.

If we consider the active stereoscopic system as shown in Figure 3, where p is the projection of P in the image plan and e the projection of C_L on the camera plan O_C, the projection of the light ray supporting the dot on the image plan is a straight line. This line is an epipolar line [18–20].

To rapidly identify a pattern point on an image we can limit the search to the epipolar lines.

For Cyclope the pattern is a regular mesh of points. For each point (j,k) of the pattern we can find the corresponding epipolar line:

$$v = a_{jk} \cdot u + b_{jk}, \tag{1}$$

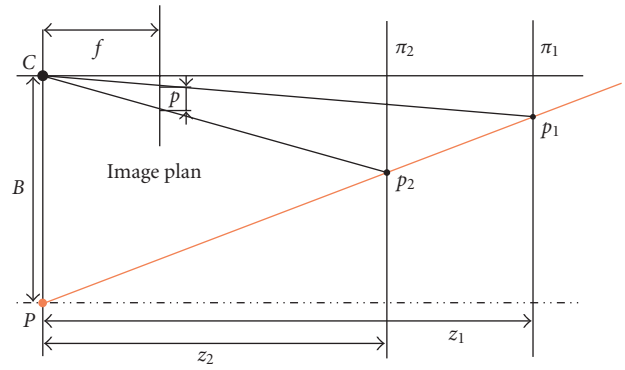


Figure 4: Spot image movement versus depth.

where (u,v) are the image coordinates and the parameters (a_{jk}, b_{jk}) are estimated through an off-line calibration process.

In addition to the epipolar lines, we can establish the relation between the position of a laser spot in the image and its distance to the stereoscopic system.

On Figure 4, we consider a laser ray (5) projected on two different plans π_1 and π_2 located, respectively, at z_1 and z_2, the trajectory d of the coordinates in the image will be constrained to the epipolar line.

By considering the two triangles CPp_1 and CPp_2, we can express d as

$$d = B\left[\left(\frac{z_1 - f}{z_1}\right) - \left(\frac{z_2 - f}{z_2}\right)\right] = Bf\frac{(z_1 - z_2)}{z_1 z_2}, \tag{2}$$

where B is the stereoscopic, f the focal length of the camera and d the distance in pixels:

$$d = \sqrt{(u_1 - u_2)^2 + (v_1 - v_2)^2}. \tag{3}$$

Given the epipolar line we can express d as a function of only one image coordinates:

$$d = \sqrt{1 + a^2}\ \ (u_1 - u_2). \tag{4}$$

From (2) and (4), we can express, for each pattern point (j,k), the depth as a hyperbolic function:

$$z = \frac{1}{\alpha_{jk} u + \beta_{jk}}, \tag{5}$$

where the α_{jk} and β_{jk} parameters are also estimated during the off-line calibration of the system [21].

We can compute the inverse of the depth z to simplify the implementation. Two operations are only needed: an addition and a multiplication. The computation of the depth of each point is independent of the others. So, all the laser spots can be computed separately allowing the parallelisation of the architecture.

3. An Energetic Approach for Multispectral Acquisition

The main problem when you design a 3D reconstruction processing for an integrated system is the limitation of

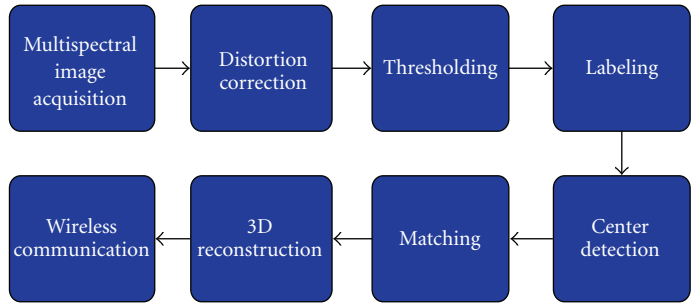


Figure 5: Acquisition and 3D reconstruction flow chart.

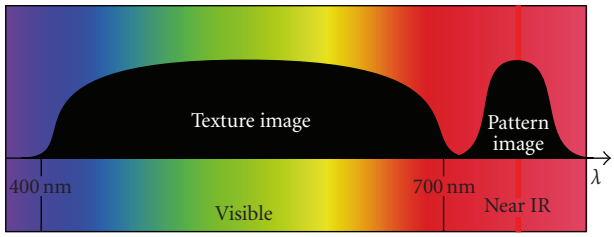


Figure 6: Multispectral image sensor.

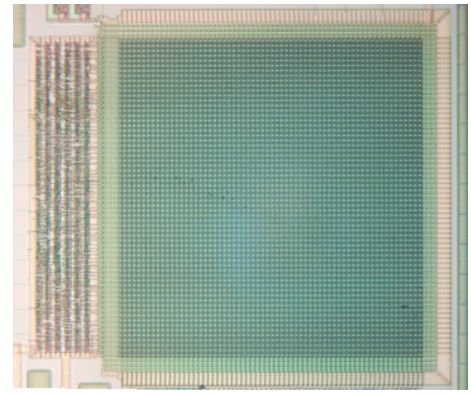


Figure 7: 64 × 64 image sensor microphotograph.

the resources. However, we can obtain a good accuracy considering hard constraints by using the following method which is shown in Figure 5:

(1) the multispectral acquisition which makes the discrimination between the pattern and the texture by an energetic method;

(2) the correction of the error coordinates due to the optical lens distortion;

(3) the processing before the 3D reconstruction as thresholding, segmentation, labelling, and the computation of the laser spot center;

(4) the computation of the matching and the third dimension;

(5) the transmission of the data with a processor core and an RF module.

The spectral response of the Silicon cuts near 1100 nm and it covers UV to near Infrared domains. This important characteristic allows defining a multispectral acquisition by grabbing on the visible band the colour texture image and, on the near infrared band, the depth information. Cyclope uses this original acquisition method, which permits to access directly at the depth information's independently from texture image processing (Figure 6).

The combination of the acquisition of the projected pattern on the infrared band, the acquisition of the texture on the visible band, and the mathematical model of the active 3D sensor makes it possible to restore the 3D textured representation of the scene. This acquisition needs to separate texture and 3D datas. For this purpose we have developed a multispectral acquisition [15]. Generally, filters are used to cut the spectral response. We used here an energetic method, which has the advantage of being generic for imagers.

To allow real-time acquisition of both pattern and texture, we have developed a first 64 × 64 pixels CMOS imager prototype in 0.6 μm for a total surface of 20 mm^2 (Figure 7). This sensor has programmable light integration and shutter time to allow dynamic change. It was designed to have large response in the visible and near infrared. This first CMOS imager prototype, which is not the subject of this article, had allowed the validation of our original energetic approach, but its small size needs to be increased to have more information. So, in our demonstrator we have used a greater CCD sensor (CIF resolution 352 × 288 pixels) to obtain normal size images and validate the 3D processing architecture.

The projector pulses periodically on the scene an energetic IR pattern. An image acquisition with a short integration time allows grabbing the image of the pattern with a background texture which appears negligible. A second image acquisition with a longer integration allows to grab the texture when the projector is off. Figure 8 shows the sequential scheduling of the images acquisition. To reach a video rate of 25 images/s this acquisition sequence must be done in less than 40 milisecond. The global acquisition time is given in (6) where T_{rst} is the reset time, T_{rd} is the time needed to read the entire image, and T_{intVI} and T_{intIR}

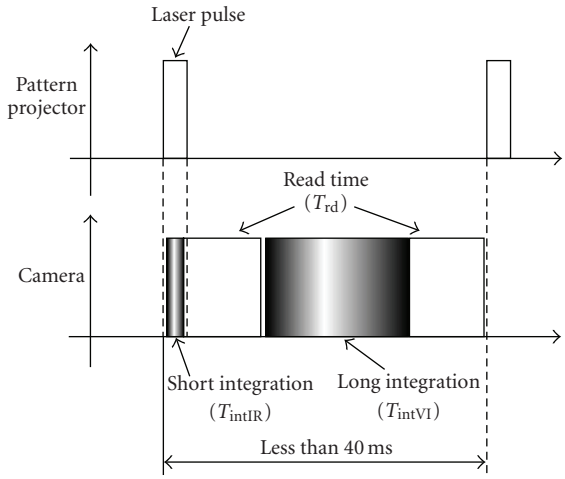


Figure 8: Acquisition sequence.

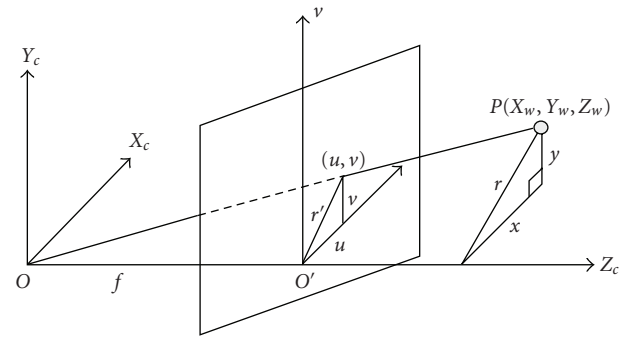


Figure 9: Pinhole camera model (X_w, Y_w, Z_w): World coordinates; (O, X_c, Y_c, Z_c): camera coordinates; (O', u, v): image plane coordinates.

are, respectively, the integration time for both visible and IR image.

$$T_{\text{total}} = 2 \cdot T_{\text{rst}} + 2 \cdot T_{\text{rd}} + T_{\text{intVI}} + T_{\text{intIR}}. \tag{6}$$

The typical values are

$$\begin{aligned} T_{\text{rst}} &= 0.5\,\text{ms}, \\ T_{\text{rd}} &= 0.5\,\text{ms}, \\ T_{\text{intVI}} &= 15\,\text{ms}, \\ T_{\text{intIR}} &= 20\,\mu\text{m}. \end{aligned} \tag{7}$$

4. Optical Distortion Correction

Generally, the lenses used in the VCE introduce large deformations on acquired images because of their weak focal [22]. This distortion is manifested in inadequate spatial relationships between pixels in the image and the corresponding points in the scene. Such change in the shape of captured object may have critical influence in medical applications, where quantitative measurements in Endoscopy depend on the position and orientation of the camera and its model. The used camera model needs to be accurate. For this reason we introduce firstly the pinhole camera model and later the correction of geometric distortion that are added to enhance it. For practical purposes two different methods are studied to implement this correction, and it is up to researchers to choose their own model depending on their required accuracy level and computational cost.

Pinhole camera model (see Figure 9) is based on the principle of linear projection where each point in the object space is projected by a straight line through the projection center into the image plane. This model can be used only as an approximation of the real camera that is actually not perfect and sustains a variety of aberration [23]. So, pinhole model is not valid when high accuracy is required like in our expected applications (Endoscopes, robotic surgery, etc.). In this case, a more comprehensive camera model must be used, taking into account the corrections for the systematically distorted image coordinates. As a result of several types of imperfections in the design and assembly of lenses composing the camera's optical system, the real projection of the point P in the image plane takes into account the error between the real image observed coordinates and the corresponding ideal (non observable) image coordinates:

$$\begin{aligned} u' &= u + \delta_u(u, v), \\ v' &= v + \delta_v(u, v), \end{aligned} \tag{8}$$

where (u, v) are the ideal nonobservable, distortion-free image coordinates, (u', v') are the corresponding real coordinates, and δ_u and δ_v are, respectively, the distortion along the u and v axes. Usually, the lens distortion consists of radial symmetric distortion, decentering distortion, affinity distortion, and nonorthogonally deformations. Several cases are presented on Figure 10.

The effective distortion can be modelled by

$$\begin{aligned} \delta_u(u, v) &= \delta_{ur} + \delta_{ud} + \delta_{up}, \\ \delta_v(u, v) &= \delta_{vr} + \delta_{vd} + \delta_{vp}, \end{aligned} \tag{9}$$

where δ_{ur} represent radial distortion [24], δ_{ud} represent decentering distortion, and δ_{up} represent thin prism distortion. Assuming that only the first- and second-order terms are sufficient to compensate the distortion, and the terms of order higher than three are negligible, we obtain a fifth-order polynomials camera model (expression 8), where (u_i, v_i) are the distorted image coordinates in pixels, and $(\tilde{u}_i, \tilde{v}_i)$ are true coordinates (undistorted):

$$\begin{aligned} u_i &= D_u S_u \Big(k_2 \tilde{u}_i^5 + 2k_2 \tilde{u}_i^3 \tilde{v}_i^2 + k_2 \tilde{u}_i \tilde{v}_i^4 + k_1 \tilde{u}_i^3 \\ &\quad + k_1 \tilde{u}_i \tilde{v}_i^2 + 3p_2 \tilde{u}_i^2 + 2p_1 \tilde{u}_i \tilde{v}_i + p_2 \tilde{v}_i^2 + \tilde{u}_i \Big) + u_0, \\ v_i &= D_v \Big(k_2 \tilde{u}_i^4 \tilde{v}_i + 2k_2 \tilde{u}_i^2 \tilde{v}_i^3 + k_1 \tilde{u}_i^2 \tilde{v}_i + k_2 \tilde{v}_i^5 \\ &\quad + k_1 \tilde{v}_i^3 + p_1 \tilde{u}_i^2 + 2p_2 \tilde{u}_i \tilde{v}_i + 3p_1 \tilde{v}_i^2 + \tilde{u}_i \Big) + v_0. \end{aligned} \tag{10}$$

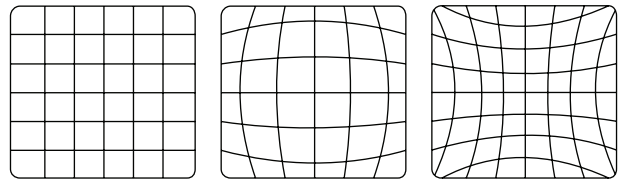


Figure 10: (a) The ideal undistorted grid. (b) Barrel distortion. (c) Pincushion distortion.

An approximation of the inverse model is done by (11):

$$\begin{aligned}
&\widetilde{u}_i \\
&= \frac{\widetilde{u}'_i + \widetilde{u}'_i\left(a_1 r_i^2 + a_2 r_i^4\right) + 2a_3\widetilde{u}'_i\widetilde{v}'_i + a_4\left(r_i^2 + 2\widetilde{u}'^2_i\right)}{a_5 r_i^2 + a_6\widetilde{u}'_i + a_7\widetilde{v}'_i + a_8)r_i^2 + 1}, \\
&\widetilde{v}_i \\
&= \frac{\widetilde{v}'_i + \widetilde{v}'_i\left(a_1 r_i^2 + a_2 r_i^4\right) + 2a_4\widetilde{u}'_i\widetilde{v}'_i + a_3\left(r_i^2 + 2\widetilde{v}'^2_i\right)}{a_5 r_i^2 + a_6\widetilde{u}'_i + a_7\widetilde{v}'_i + a_8)r_i^2 + 1},
\end{aligned} \tag{11}$$

where

$$\begin{aligned}
\widetilde{u}'_i &= \frac{u_i - u_0}{D_{uSU}}, \\
\widetilde{v}'_i &= \frac{v_i - v_0}{D_v}, \\
r_i^2 &= \sqrt{\widetilde{v}'^2_i + \widetilde{v}'^2_i}.
\end{aligned} \tag{12}$$

The unknown parameters $a_1, \ldots, a_8$ are solved using direct least mean-squares fitting [25] in the off-line calibration process.

4.1. Off-Line Lens Calibration. There are many proposed methods that can be used to estimate intrinsic camera and lens distortion parameters, and there are also methods that produce only a subset of the parameter estimates. We chose a traditional calibration method based on observing a planar checkerboard in front of our system at different poses and positions (see Figure 11) to solve the equations of unknown parameters (11). The results of the calibration procedure are presented in Table 1.

4.2. Hardware Implementation. After the computation of parameters in (11) through an off-line calibration process, we used them to correct the distortion of each frame. With the input frame captured by the camera denoted as the source image and the corrected output as the target image, the task of correcting the source distorted image can be defined as follows: for every pixel location in the target image, compute its corresponding pixel location in the source image. Two implementation techniques of distortion correction have been compared:

Direct Computation. Calculate the image coordinates through evaluating the polynomials to determine intensity values for each pixel.

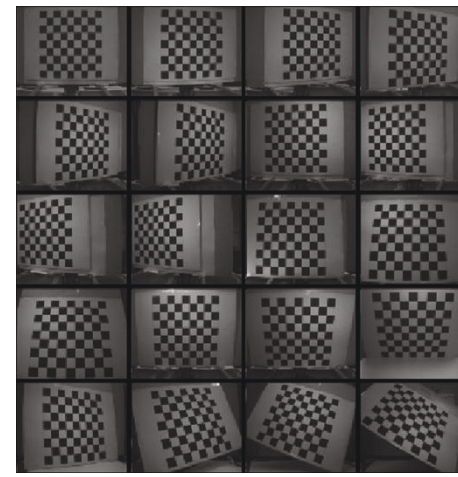


Figure 11: Different checkerboard positions used for calibration procedure.

Table 1: Calibration results.

Parameter	Value	Error
u_0 (pixels)	178.04	1.28
v_0 (pixels)	144.25	1.34
fD_{uSU} (pixels)	444.99	1.21
fD_v (pixels)	486.39	1.37
a_1	−0.3091	0.0098
a_2	−0.0033	0.0031
a_3	0.0004	0.0001
a_4	0.0014	0.0004
a_5	0.0021	0.0002
a_6	0.0002	0.0001
a_7	0.0024	0.0005
a_8	0.0011	0.0002

Lookup Table. Calculate the image coordinates through evaluating the polynomials correction in advance, storing them in a lookup table which is referenced at run-time. All parameters needed for LUT generation are known beforehand; therefore for our system, the LUT is computed only once and off-line.

However, since the source pixel location can be a real number, using it to compute the actual pixel values of the target image requires some form of pixel interpolation. For this purpose we have used the nearest neighbour

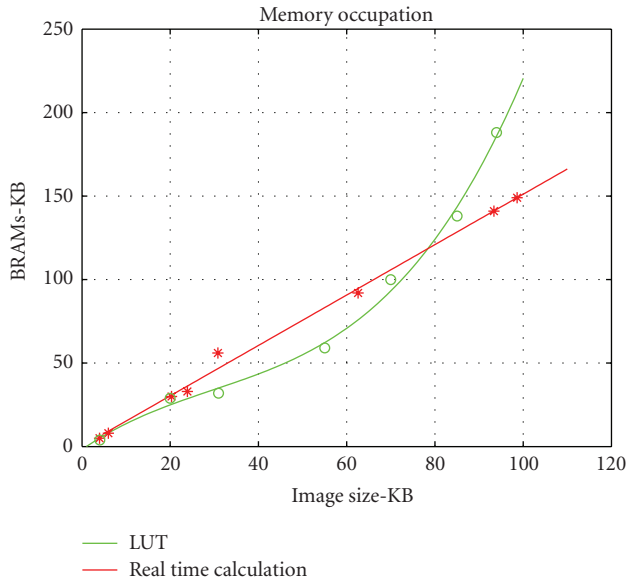


Figure 12: Block Memory occupation for both direct computation and LUT-based approaches.

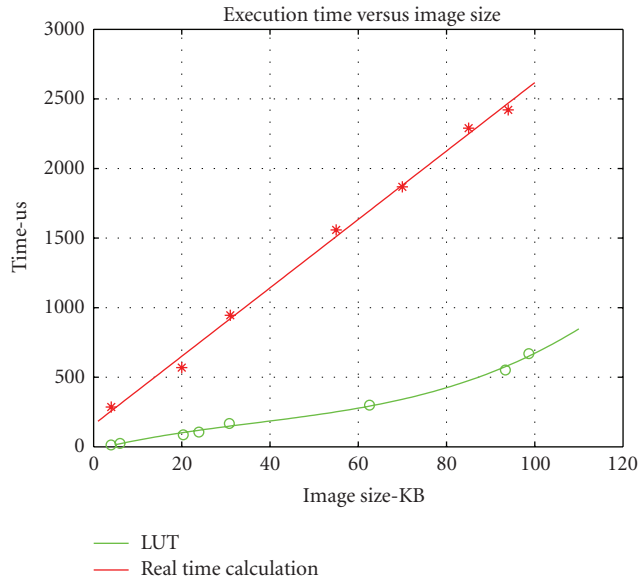


Figure 13: Execution time for both direct computation and LUT-based approaches.

Table 2: Area and Clock characteristics of two approaches.

Implementation	Area (%)	Clock (MHz)
Direct Computation	58	10
Look Up Table	6	24

interpolation approach that means that the pixel value closest to the predicted coordinates is assigned to the target coordinates. This choice is reasonable because it is a simple and fast method for computation, and visible image artefacts have no subject with our system.

Performance results of these two techniques are presented in terms of (i) execution time and (ii) FPGA logic resource requirements.

The proposed architectures described above have been described in VHDL in a fixed point fashion, implemented on a Xilinx Virtex II FPGA device and simulated using industry reference simulator (ModelSim). The pixel values of both the input distorted and the output corrected images use an 8-bit word length integer number. The coordinates use an 18-bit word length.

The results are presented in Figures 12 and 13 and Table 2.

The execution time for the direct computation implementation is comparatively very slow. This is due to the fact that the direct computation approach consumes a much greater amount of logic resources than the Look-up Table approach. Moreover the slow clock cycle (10 MHz) could be increased by splitting the complex arithmetic logic into several smaller stages. The significant difference between these two approaches is that the direct computation approach requires more computation time and arithmetic operations, while the LUT approach requires more memory accesses and more RAM Blocks occupation. Regarding latency, both approaches can be executed with respect to real-time constraint of video cadence (25 frames per second). Depending on the applications, the best compromise between time and resources must be chosen by the user. For our application, arithmetic operations are intensively needed for later stages in the preprocessing block, while memory blocks are available; so we chose to use the LUT approach to benefit in time and resources.

5. Thresholding and Labelling

After lens distortion correction, the laser spots projected must be extracted from the gray level image for delivering a 3D representation of the scene. Laser spots on the image appear with variable sizes (depending on the absorption of the surface and the projection angle). At this level, a preprocessing block has been developed and hardware implemented to make an adaptive thresholding in order to give a binary image and a labelling to classify each laser spot to compute later their center.

5.1. Thresholding Algorithm. Several methods exist from a static threshold value defined by user up to dynamic algorithm as Otsu method [26].

We have chosen to develop a new approach less complex than Otsu or others well-known dynamic methods in order to reduce the processing time [27]. The simple method is described in (Figure 14):

(i) building the histogram of grey-level image,

(ii) finding the first maxima of the Gaussian corresponding to the Background; compute its mean μ and standard deviation σ,

(iii) calculating the threshold value with (13):

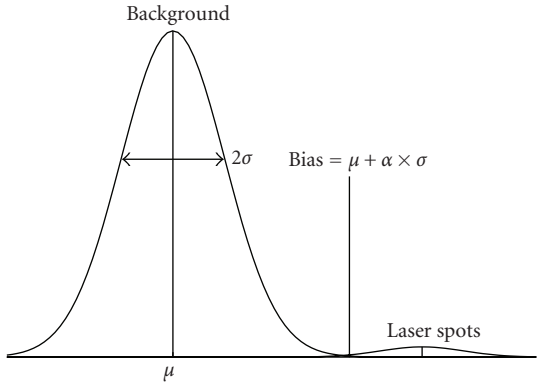


Figure 14: Method developed in Cyclope.

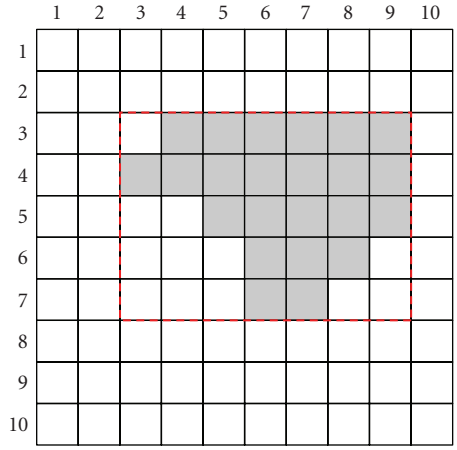


Figure 15: Smallest rectangle containing active pixel.

$$\text{Threshold} = \sigma \cdot \alpha + \mu, \tag{13}$$

where α is an arbitrary constant. A parallel architecture of processing has been designed to compute the threshold and to give a binary image. Full features of this implementation are given in [28].

5.2. Labeling. After this first stage of extraction of spot laser from the background, it is necessary to classify each laser spot in order to compute separately their center. Several methods have been developed in the past. We chose to use a classical two passes component connected labeling algorithms with an 8-connectivity. We designed a specific optimized Intellectual Property in VHDL. This intellectual property uses fixed point number.

6. Computation of Spots Centers

The threshold and labeling processes applied to the captured image allow us to determine the area of each spot (number of pixels). The coordinates of center of these spots could be calculated as follows:

$$u_{gI} = \frac{\sum_{i \in I} u_i}{N_I}, \qquad v_{gI} = \frac{\sum_{i \in I} v_i}{N_I}, \tag{14}$$

where u_{gI} and v_{gI} and the abscissa and ordinate of Ith spot center. u_i and v_i are the coordinates of pixels constructing the spot. N_I is the number of pixels of Ith spot (area in pixels).

To obtain an accuracy 3D reconstruction, we need to compute the spots centers with higher possible precision without increasing the total computing time to satisfy the real-time constraint. The hardest step in center detection part is the division operations A/B in (14). Several methods exist to solve this problem.

6.1. Implementation of a Hardware Divider. The simplest method is the use of hardware divider but they are computationally expensive and consume a considerable amount of resources. This not acceptable for a real-time embedded systems. Some other techniques are used to compute the center of laser spots avoiding the use of hardware dividers.

6.2. Approximation Method. Some studies suggest approximation methods to avoid implementation of hardware dividers. Such methods like that implemented in [29] replace the active pixels by the smallest rectangle containing this region and then replace the usual division by simple shifting (division by 2):

$$u^*_{gI} = \frac{\text{Max}(u_i) + \text{Min}(u_i)}{2}, \qquad v^*_{gI} = \frac{\text{Max}(v_i) + \text{Min}(v_i)}{2}. \tag{15}$$

This approach is approximated in (15), where (u_i, v_i) are the active pixel coordinates, and (u^*_{gI}, v^*_{gI}) are the approximated coordinates of the spot center.

The determination of rectangle limits needs two times scanning of the image, detecting in every scanning step, respectively, the minimum and maximum of pixels coordinates. For each spot, we should compare the coordinates of every pixel by last registered minimum and maximum to assign new values to U_m, U_M, V_m, and V_M. (m: Minimum; M: maximum). While N_p is the average area of spots (number of pixels), we can estimate the number of operations needed to calculate the center of each spot by $4N_p + 6$. And in global, $N_{\text{op}} \approx 25 * N * (4N_p + 6)$ operations are needed to calculate the centers of N spots (video-cadence of 25 fps). Such approximation is simple and easy to use but still needs considerable time to be calculated. Beside, the error is not negligible. The error in such method is nearly 0.22 pixel, and the maximum error is more than 0.5 pixel [29]. Taking the spot of Figure 15 as an example of inaccuracy of such a method, the real center position of these pixels is (4.47; 6.51). But when applying this approximation method, the center position will be (5; 6). This inaccuracy will result mismatching problem that affects the measurement result when reconstructing object.

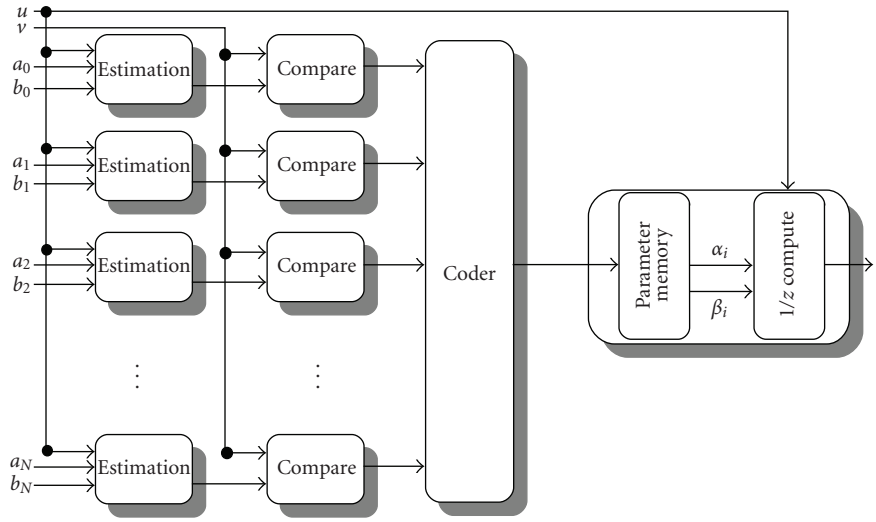


Figure 16: 3D unit.

6.3. Our Method. The area of each spot (number of pixels) is always a positive integer, while its value is limited in a predeterminate interval $[N_{\min}, N_{\max}]$, where $N_{\min}$ and $N_{\max}$ are, respectively, the minimum and maximum areas of laser spot in the image. The spot areas depend on object illumination, distance between object and camera, and the angle of view of the scene. Our method consists in a memorisation of $1/N$ where N represent the spot pixel number and can take value in $[1, N_{\text{limit}}]$. N_{limit} represent the maximum considered size, in pixels, of a spot.

In this case we need only to compute a multiplication, that is resume here:

$$u_{gI} = (u_1 + u_2 + \cdots + u_I) * \frac{1}{N_I},$$
$$v_{gI} = (v_1 + v_2 + \cdots + v_I) * \frac{1}{N_I}. \tag{16}$$

The implementation of such a filter is very easy, regarding that the most of DSP functions are provided for earlier FPGAs. For example, Virtex-II architecture [30] provides an 18×18 bits Multiplier with a latency of about 4.87 ns at 205 MHz and optimised for high-speed operations. Additionally, the power consumption is lower compared to a slice implementation of an 18-bit by 18-bit multiplier [31]. For N luminous spots source, the number of operations needed to compute the centers coordinates is $N_{\text{op}} \approx 25 * N * N_P$, and N_p is the average area of spots. When implementing our approach to Virtex II Pro FPGA (XC2VP30), it was clear that we gain in execution time and size. Comparison of different implementation approaches is described in the next section.

7. Matching Algorithm

The set of parameters for the epipolar and depth models are used during run time to make point matching (identify the original position of a pattern point from its image) and calculate the depth using the coordinates of each laser spot center.

For this purpose we have developed a parallel architecture visible in Figure 16, described in detail in [32].

Starting from the point abscissa (u) we calculate its estimated ordinate ($\tilde{v}$) if it belongs to a epipolar line. We compare this estimation with the true ordinate (v).

These operations are made for all the epipolar line simultaneously. After thresholding the encoder returns the index of the corresponding epipolar line.

The next step is to calculate the z coordinate from the u coordinate and the appropriate depth model parameters (α, β)

These computation blocs are synchronous and pipelined, allowing, thus, high processing rates.

7.1. Estimation Bloc. In this bloc the estimated ordinate is calculated $\tilde{v} = a \cdot u + b$. The (a, b) parameters are loaded from memory.

7.2. Comparison Bloc. In this bloc the absolute value of the difference between the ordinate v and its estimation $\tilde{v}$ is calculated. This difference is then thresholded.

The thresholding avoids a resource consuming sort stage. The threshold was a priori chosen as half the minimum distance between two consecutive epipolar lines. The threshold can be adjusted for each comparison bloc.

This bloc returns a "1" result if the distance is underneath the threshold.

7.3. Encoding Bloc. If the comparison blocs return a unique "1" result, then the encoder returns the corresponding epipolar line index.

If no comparison bloc returns a "true" result, the point is irrelevant and considered as picture noise.

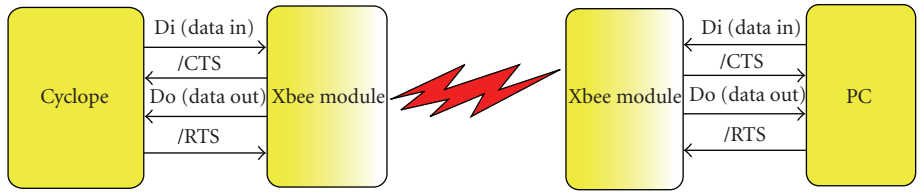


Figure 17: Wireless communication.

If more than one comparison blocs returns "1", then we consider that we have a correspondence error and a flag is set.

The selected index is then carried to the next stage where the z coordinate is calculated. It allows the selection of the right parameters to the depth model.

We compute $(1/z)$, rather than z as we said earlier, to have a simpler computation unit. This computation bloc is then identical to the estimation bloc.

8. Wireless Communication

Finally, after computation, the 3D coordinates of the laser dots accompanied by the image of texture are sent to an external reader. So, Cyclope is equipped with a block of wireless communication which allows us to transmit the image of texture, the coordinates 3D of the centers of the spots laser, and even to remotely reconfigure the digital processing architecture (an Over The Air FPGA). While attending the IEEE802.15 Body Area Network standard [33], the frequency assigned for implanted device RF communication is around 403 MHz and referred to as the MICS (Medical Implant Communication System) band due to essentially three reasons:

(i) a small antenna,

(ii) a minimum losses environment which allows to design low-power transmitter,

(iii) a free band without causing interference to other users of the electromagnetic radio spectrum [34].

In order to make rapidly a wireless communication of our prototype, we chose to use Zigbee module at 2.45 GHz available on the market contrary to modules MCIS. We are self-assured that later frequency is not usable for the communication between the implant and an external reader, due to the electromagnetic losses of the human body. Two Xbee-pro modules from the Digi Corporation have been used. One for the demonstrator and the second plugged on a PC host where a human machine interface has been designed to visualise in real-time the 3D textured reconstruction of the scene.

Communication between wireless module and the FPGA circuit is performed by a standard UART protocol. this principle is shown on Figure 17. To make this communication we integrated a Microblaze softcore processor with UART functionality. The Softcore recovers all the data stored in memory (texture and 3D coordinates) and sends them to the wireless module.

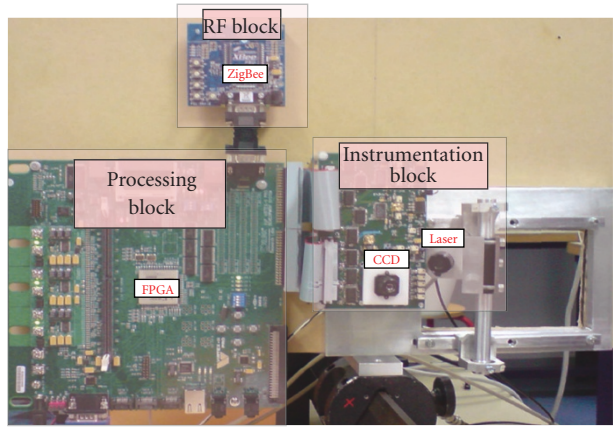


Figure 18: Demonstrator

9. Demonstrator, Testbench and Results

9.1. Experimental Demonstrator. To demonstrate the feasibility of our system, a large-scale demonstrator has been realised. It uses an FPGA prototyping board based on a Xilinx Virtex2Pro, a pulsed IR LASER projector [35] coupled with a diffraction network that generates a 49-dot pattern and a CCD imager.

Figure 18 represents the experimental set. It is composed of a standard 3 mm lens, the CCD camera with an external 8 bits DAC, a projector IR pattern, and a Virtex2pro prototyping board.

FPGA is used mainly for computation unit but also to control image acquisition, laser synchronisation, analog-to-digital conversion, and image storage and displays the result through a VGA interface.

Figure 19 shows the principal parts of the control and storage architecture as set in the FPGA. Five parts have been designed:

(i) a global sequencer to control the entire process,

(ii) a reset and integration time configuration unit,

(iii) a VGA synchronisation interface,

(iv) a dual port memory to store the images and to allow asynchronous acquisition and display operations,

(v) a wirless communication module based on the ZigBee protocol.

A separated pulsed IR projector has been added to the system to demonstrate the system functionality.

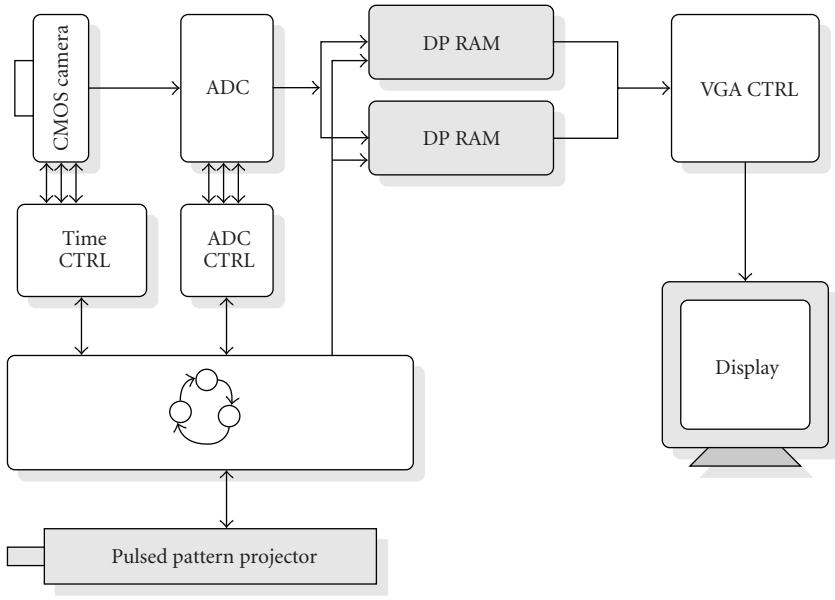


Figure 19: Implementation of the control and storage architecture.

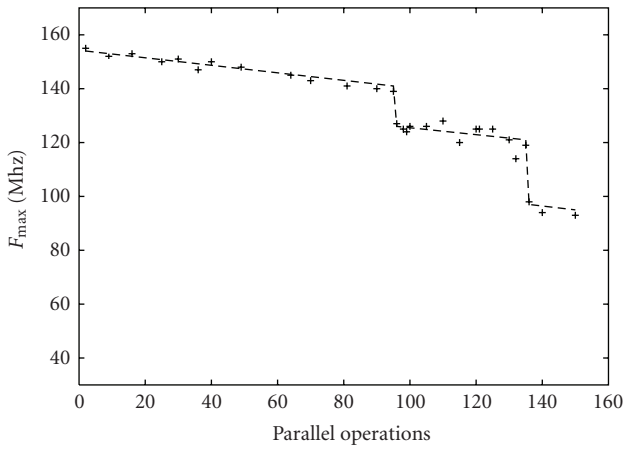
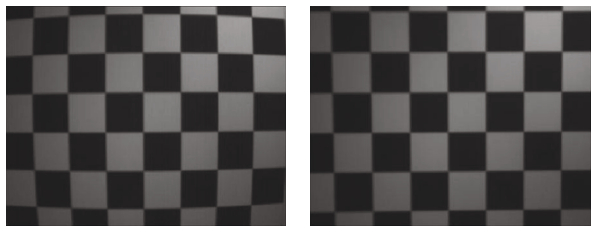


Figure 20: FPGA working frequency evolution.

Table 3: Performances of the distortion correction.

Slices	1795 (13%)
Latency	11.43 ms
Error	< 0.01 pixels



(a) Image without correction (b) Image with correction

Figure 21: (a) Checkerboard image before distortion correction. (b) Checkerboard image after correction.

The computation unit was described in VHDL and implemented on FPGA XilinX VirtexIIpro (XC2VP30) with 30816 logic cells and 136 hardware multipliers [31]. The synthesis and placement were achieved for 49 parallel processing elements. We use here 28% of the LUTs and 50 hardware multipliers, for a working frequency of 148 Mhz.

9.2. Architecture Performance. To estimate the evolution of the architecture performances, we have used a generic description and repeat the synthesis and placement for different pattern sizes (number of parallel operations). Figure 20 shows that in every case our architecture mapped on an FPGA can work at least at almost 90 Mhz and then obtain a real time constraint of 40 milliseconds.

9.3. Error Estimation of the Optical Correction. The implementation results of distortion correction method are summarised in Table 3. In this table we have implemented the correction model only to the active light spots. However, Figure 21 present an image before and after our lens distortion correction.

Regarding size and latency, it is clear that the results are suitable for our application.

Comparing our used method to compute the spots centers with two other methods (see Table 4), it is clear that our approach has higher accuracy and smaller size than approximation method. Since it has nearly the same accuracy as method using hardware divider, it still uses less resources.

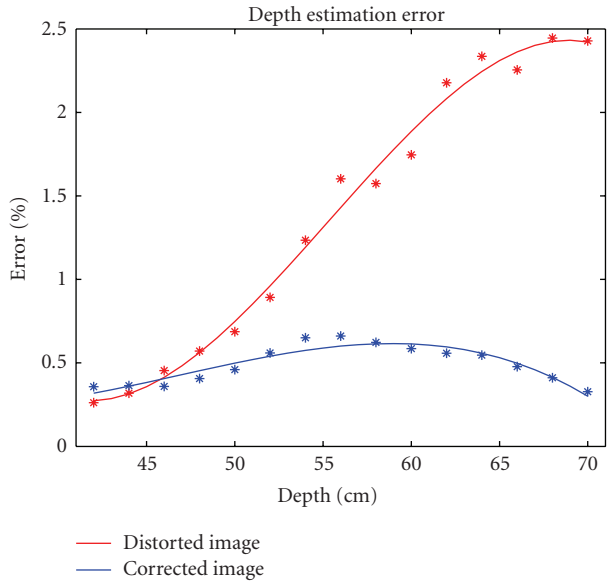


Figure 22: Error comparison before and after applying distortion correction and centers recomputing.

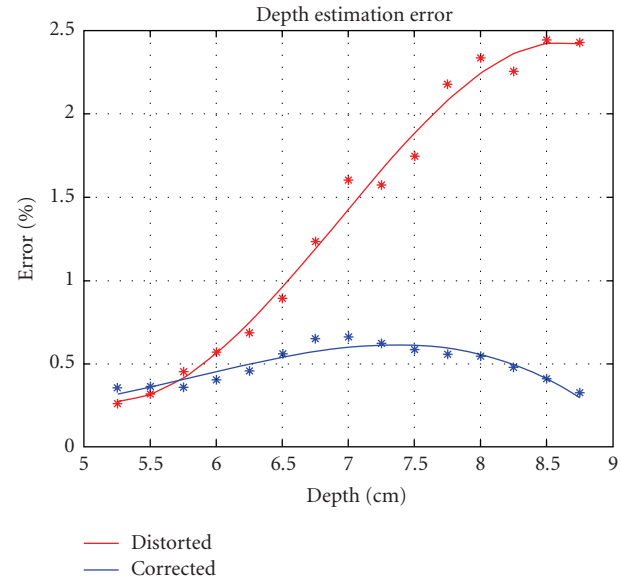


Figure 23: Error comparison before and after applying distortion correction and centers recomputing after scaling for integration.

Regarding latency, the results of all three approaches respect real time constraint of video cadence (25 frames per second). Comparing many measures on the depth estimation before and after the implementation of our improvements, the results indicate that the precision of the system increased, so that the residual error is reduced about 33% (Figure 22).

These results were interpolated with a scale factor to measure the error lens in the case of integration inside a video capsule, and the results can be shown in Figure 23. This scaling was calculated with a distance between the laser projector and the imager of 1 cm. It is the maximal distance that can be considered for endoscopy. This distance corresponds to the diameter of the PillCam video capsule. We can show that the correction of the distortion produced by the lens increases the accuracy of our sensor.

9.4. Error Estimation of the 3D Reconstruction. In order to validate our reconstruction architecture, we have compared the results obtained with the synthesised IP (Table 5) and those obtained from a floating point mathematical model which was already validated by experimentation. As we can see, the calculation error margin is relatively weak in comparison with the distance variations and shows that our approach to translate a complex mathematical model into a digital processing for embedded system is valid.

Table 6 shows the error of reconstruction for different distances and sizes of the stereoscopic base. We can see that for a base of 5 mm we are able to have a 3D reconstruction with an error above to 4% at a distance of 10 cm. This precision is perfectly enough in the context of the human body exploration and an integration of a stereoscopic base with a such size is relatively simple.

Table 4: Centers computation performances.

Method	Slices	Latency (μs)	Error (pixel)
Approximation	287	4.7	0.21
Hardware divider	1804	1.77	0.0078
Our approach	272	2.34	0.015

Table 5: Results Validation.

Coordinates couples abscise/ordinate (pixel)	Model results (meter)	IP results (meter)
401/450	1.57044	1.57342
357/448	1.57329	1.57349
402/404	1.57223	1.57176
569/387	1.22065	1.21734
446/419	1.11946	1.11989
478/319	1.07410	1.07623
424/315	1.04655	1.04676
375/267	1.03283	1.03297
420/177	1.03316	1.03082

Table 6: Precision versus the size of the stereoscopic base.

Base of 0.5 cm		Base of 1.5 cm	
Distance (cm)	Error (%)	Distance (cm)	Error (%)
5	1.8	5	0.61
10	3.54	10	1.21
50	15.52	50	5.77
100	26.87	100	10.91

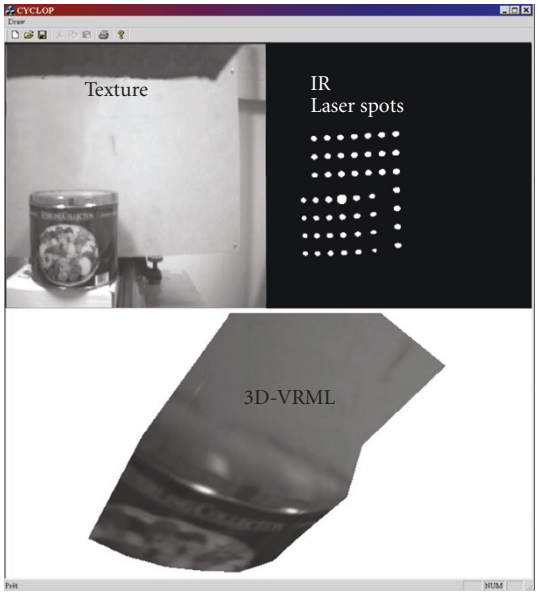


Figure 24: Visualisation of the results by our application.

9.5. Example of Reconstruction. We have used the calibration results to reconstruct the volume of an object (a 20 cm diameter cylinder). The pattern was projected on the scene and the snapshots were taken.

The pattern points were extracted and associated to laser beams using the epipolar constraint. The depth of each point was then calculated using the appropriate model. The texture image was mapped on the reconstructed object and rendered in an VRML player.

We have created an application written in C++ for visualising the Cyclope's results (Figure 24). The application gets the textural information and the barycenter spatial position of the 49 infrared laser spots from a wireless communication module. After the result reception, it draws the texture of three binary maps representing the location of the 49 barycenters on a 3D coordinates system (XY, ZX, and ZY).

The recapitulation of the hardware requirement is presented Table 7. We can observe that the design is small and if we make an equivalence in logics gates, it should be integrated in a small area chip like IGLOO AGL 1000 device from Actel. Such a device has a size of $10 \times 10\,\text{mm}^2$ and its core can be integrated in a VCE which has a diameter around 1 cm. At this moment, we did not make implementation on this last platform. It is a feasibility study but the first results prove that this solution is valid if we consider the needed ressources.

We present also an estimation of the energetic consumption which was realised with two tools. This estimation is visible in Table 8. The first tool is XPE (Xilinx Power Estimation) from Xilinx to evaluate the power consumption of a Virtex, and the second is IGLOOpowercalculator from Actel to evaluate the power consumption of a low power consumption FPGA.

Table 7: Recapitulation of the performances.

Architecture	Clb slices	Latches	LUT	RAM
Camera	309	337	618	4
Optical correction*	92/94	8/8	176/190	32/56
Thresholding	107	192	214	1
Labelling	114	102	227	0
Matching	1932	3025	3864	0
Communication	170	157	277	3
Total used*	2323/2325	3821	1555/1569	40/64
Total free	13693	29060	27392	136

*Direct computation/Look up table.

Table 8: Processing block power consumption estimation.

Device	Power consumption	Duration	
		1 battery	3 battery
Virtex	1133 mW	29 min	1h 26 min
IGLOO	128,4 mW	4 hours	12 hours

These two tools use the processing frequency, the number of logic cells, the number of D flip-flop, and the amount of memory of the design to estimate the power consumption. To realise our estimation, we use the results summarised in Table 7. Our estimation is made with an activity rate of 50% that is the worst case.

To validate the power consumption estimation in an embedded context, we consider that a 3V-CR1220 battery (3V-CR1220 is a 3 Volt battery, its diameter is of 1.2 cm, and its thickness is of 2 mm) which has a maximum of 180 mAh power consumption, that is to say an ideal power of 540 mWh. This battery is fully compatible with a VCE like the Pillcam from Given Imaging.

As we can see, the integration of a Virtex in a VCE is impossible because of the SRAM memory that consumes too much energy. If we consider the IGLOO technology based on flash memory, we can observe that its power consumption is compatible with a VCE. Such technology permits four hours of autonomy with only one battery, and twelve hours of autonomy if we used three 3V-CR1220 in the VCE. This result is encouraging because at this time the mean duration of an examination is ten hours.

10. Conclusion and Perspectives

We have presented in this paper Cyclope, a sensor designed to be a 3D video capsule.

we have explained a method to acquire the images at a 25-frame/s video rate with a discrimination between the texture and the projected pattern. This method uses an energetic approach, a pulsed projector, and an original 64×64 CMOS image sensor with programmable integration time. Multiple images are taken with different integration times to obtain an image of the pattern which is more energetic than the background texture. Our CMOS imager validates this method.

Also we present a 3D reconstruction processing that allows a precise and real-time reconstruction. This processing which is specifically designed for an integrated sensor and its integration in an FPGA-like device has a low power consumption compatible with a VCE examination.

The method was tested on a large scale demonstrator using an FPGA prototyping board and a 352×288 pixels CCD sensor. The results show that it is possible to integrate a stereoscopic base which is designed for a integrated sensor and to keep a good precision for a human body exploration.

The next step to this work is the chip level integration of both the image sensor and the pattern projector. Evaluate the power consumption of the pulsed laser projector considering the optical efficiency of the diffraction head.

The presented version of Cyclope is the first step toward the final goal of the project. After this, the goal is to realise a real-time pattern recognition with processing-like support vector machine or neuronal network. The final issue of Cyclope is to be a real smart sensor that can realize a part of a diagnosis inside the body and then increase its fiability.

References

[1] G. Iddan, G. Meron, A. Glukhovsky, and P. Swain, "Wireless capsule endoscopy," *Nature*, vol. 405, no. 6785, pp. 417–418, 2000.

[2] J.-F. Rey, K. Kuznetsov, and E. Vazquez-Ballesteros, "Olympus capsule endoscope for small and large bowel exploration," *Gastrointestinal Endoscopy*, vol. 63, no. 5, p. AB176, 2006.

[3] M. Gay, et al., "La vidéo capsule endoscopique: qu'en attendre?" CISMEF, http://www.churouen.fr/ssf/equip/capsules-videoendoscopiques.html.

[4] T. Graba, B. Granado, O. Romain, T. Ea, A. Pinna, and P. Garda, "Cyclope: an integrated real-time 3d image sensor," in *Proceedings of the 19th International Conference on Design of Circuits and Integrated Systems*, 2004.

[5] F. Marzani, Y. Voisin, L. L. Y. Voon, and A. Diou, "Active sterovision system: a fast and easy calibration method," in *Proceedings of the 6th International Conference on Control Automation, Robotics and Vision (ICARCV '00)*, 2000.

[6] W. Li, F. Boochs, F. Marzani, and Y. Voisin, "Iterative 3d surface reconstruction with adaptive pattern projection," in *Proceedings of the 6th IASTED International Conference on Visualization, Imaging, and Image Processing (VIIP '06)*, pp. 336–341, August 2006.

[7] P. Lavoie, D. Ionescu, and E. Petriu, "A high precision 3d object reconstruction method using a color coded grid and nurbs," in *Proceedings of the International Conference on Image Analysis and Processing*, 1999.

[8] Y. Oike, H. Shintaku, S. Takayama, M. Ikeda, and K. Asada, "Real-time and high resolution 3-d imaging system using light-section method and smart CMOS sensor," in *Proceedings of the IEEE International Conference on Sensors (SENSORS '03)*, vol. 2, pp. 502–507, October 2003.

[9] A. Ullrich, N. Studnicka, J. Riegl, and S. Orlandini, "Long-range highperformance time-of-flight-based 3d imaging sensors," in *Proceedings of the International Symposium on 3D Data Processing Visualization and Transmission*, 2002.

[10] A. Mansouri, A. Lathuiliére, F. S. Marzani, Y. Voisin, and P. Gouton, "Toward a 3d multispectral scanner: an application to multimedia," *IEEE Multimedia*, vol. 14, no. 1, pp. 40–47, 2007.

[11] F. Bernardini and H. Rushmeier, "The 3d model acquisition pipeline," *Computer Graphics Forum*, vol. 21, no. 2, pp. 149–172, 2002.

[12] S. Zhang, "Recent progresses on real-time 3d shape measurement using digital fringe projection techniques," *Optics and Lasers in Engineering*, vol. 48, no. 2, pp. 149–158, 2010.

[13] F. W. Depiero and M. M. Triverdi, "3d computer vision using structured light: design, calibration, and implementation issues," *Journal of Advances in Computers*, pp. 243–278, 1996.

[14] E. E. Hemayed, M. T. Ahmed, and A. A. Farag, "CardEye: a 3d trinocular active vision system," in *Proceedings of the IEEE Conference on Intelligent Transportation Systems (ITSC '00)*, pp. 398–403, Dearborn, Mich, USA, October 2000.

[15] A. Kolar, T. Graba, A. Pinna, O. Romain, B. Granado, and E. Belhaire, "Smart Bi-spectral image sensor for 3d vision," in *Proceedings of the 6th IEEE Conference on SENSORS (IEEE SENSORS '07)*, pp. 577–580, Atlanta, Ga, USA, October 2007.

[16] B. Gyselinckx, C. Van Hoof, J. Ryckaert, R. F. Yazicioglu, P. Fiorini, and V. Leonov, "Human++: autonomous wireless sensors for body area networks," in *Proceedings of the IEEE Custom Integrated Circuits Conference*, pp. 12–18, 2005.

[17] B. Warneke, M. Last, B. Liebowitz, and K. S. J. Pister, "Smart dust: communicating with a cubic-millimeter computer," *Computer*, vol. 34, no. 1, pp. 44–51, 2001.

[18] R. Horaud and O. Monga, *Vision par Ordinateur*, chapter 5, Hermès, 1995.

[19] O. Faugeras, *Three-Dimensional Computer Vision, a Geometric Viewpoint*, MIT Press, Cambridge, Mass, USA, 1993.

[20] J. Batlle, E. Mouaddib, and J. Salvi, "Recent progress in coded structured light as a technique to solve the correspondence problem: a survey," *Pattern Recognition*, vol. 31, no. 7, pp. 963–982, 1998.

[21] S. Woo, A. Dipanda, F. Marzani, and Y. Voisin, "Determination of an optimal configuration for a direct correspondence in an active stereovision system," in *Proceedings of the IASTED International Conference on Visualization, Imaging, and Image Processing*, 2002.

[22] O.-Y. Mang, S.-W. Huang, Y.-L. Chen, H.-H. Lee, and P.-K. Weng, "Design of wide-angle lenses for wireless capsule endoscopes," in *Optical Engineering*, vol. 46, October 2007.

[23] J. Heikkilä and O. Silvén, "A four-step camera calibration procedure with implicit image correction," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1106–1112, San Juan, Puerto Rico, USA, 1997.

[24] K. Hwang and M. G. Kang, "Correction of lens distortion using point correspondence," in *Proceedings of the IEEE Region 10 Conference (TENCON '99)*, vol. 1, pp. 690–693, 1999.

[25] J. Heikkilä, *Accurate camera calibration and feature based 3-D reconstruction from monocular image sequences*, Ph.D. dissertation, University of Oulu, Oulu, Finland, 1997.

[26] N. Otsu, "A threshold selection method from gray level histogram," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[27] J. N. Kapur, P. K. Sahoo, and A. K. C. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram," *Computer Vision, Graphics, & Image Processing*, vol. 29, no. 3, pp. 273–285, 1985.

[28] D. Faura, T. Graba, S. Viateur, O. Romain, B. Granado, and P. Garda, "Seuillage dynamique temps réel dans un système embarqué," in *Proceedings of the 21ème Colloque du Groupe de Recherche et d'Étude du Traitement du Signal et des l'Image (GRETSI '07)*, 2007.

[29] T. Graba, *Etude d'une architecture de traitement pour un capteur intégré de vision 3d*, Ph.D. dissertation, Université Pierre et Marie Curie, 2006.
[30] M. Adhiwiyogo, "Optimal pipelining of the I/O ports of the virtex-II multiplier," XAPP636, vol. 1.4, June 2004.
[31] Xilinx, "Virtex-II Pro and Virtex-II Pro Platform FPGA: Complete Data Sheet," October 2005.
[32] A. Kolar, T. Graba, A. Pinna, O. Romain, B. Granado, and T. Ea, "A digital processing architecture for 3d reconstruction," in *Proceedings of the International Workshop on Computer Architecture for Machine Perception and Sensing (CAMPS '06)*, pp. 172–176, Montreal, Canada, August 2006.
[33] ieee802, http://www.ieee802.org/15/pub/TG6.html.
[34] M. R. Yuce, S. W. P. Ng, N. L. Myo, J. Y. Khan, and W. Liu, "Wireless body sensor network using medical implant band," *Journal of Medical Systems*, vol. 31, no. 6, pp. 467–474, 2007.
[35] Laser2000, http://www.laser2000.fr/index.php?id=368949&L=2.