A Low-Power Stereo Vision System Based on a Custom CMOS Imager with Positional Data Coding

Nicola Cottini, Marco De Nicola, Massimo Gottardi
Fondazione Bruno Kessler
38050 Povo, Trento, I, cottini@fbk.eu

Roberto Manduchi
Dept. of Computer Engineering University of California
Santa Cruz, CA 95064, USA, manduchi@soe.ucsc.edu

Abstract — **A new stereo vision system is presented, which is based on a low-power CMOS binary vision sensor, with embedded spatial contrast extraction and output data compression, through positional data coding. A novel disparity algorithm adapted to the output of the sensors has been developed. The algorithm has been simulated and tested on a real application. The presented system is targeted towards event-based applications, where information is only provided under significant changes in the scene. Due to the custom address-driven data coding of the sensors, a novel approach for image processing has been adopted, which is more efficient with respect to those adopted in standard raster-scan imagers. The system has been tested in a people counting application, installing the prototype on the ceiling, at around 3m height. In this way, the system is able to discriminate people heights with an accuracy of about 5cm.**

Keywords — CMOS vision sensors, low-power sensors, stereo-vision, image processing.

I. INTRODUCTION

Advances in standard CMOS Active Pixel Sensors (APS), thanks to the availability of deep sub-micron technology, have brought the development of high-performance imagers, with reduced pixel pitch, large spatial resolution and high miniaturization level. With the constant increase of spatial resolution, following the technology scale down, the imager has to dispatch a large amount of data at high speed to guarantee video frame-rate. 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. Ad hoc vision sensors are better targeted to salient feature extraction from the scene with reduced energy budget, rather than reproducing the exact photographic still picture scene. Recently, some efforts have been devoted to the development of micropower imagers [1-4]. Although their advanced performance in power consumption, those sensors are not properly targeted to minimize the energy budget of the overall vision system. In fact, most of them deliver data in a standard raster-scan mode, without any kind of data pre-filtering, which would make the job of the image processor much lighter and less power hungry. In this work, we present a stereo-vision system, based on a custom 128x64 pixel CMOS imager [1], performing on-pixel contrast extraction and binarization with a power consumption of about 60µW at 30frame/s. Moreover, 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. This turns into a drastic data reduction at the sensor interface, which is typically 5-10% of the total imager resolution. In Active mode, only a small amount of data need to be processed, with a reduced amount of computation. Here the main issue is that data are organized in a sparse matrix, thus requiring non conventional image processing, which are typically based on kernel concept and spatial neighborhood correlation. For the sake of simplicity, Fig. 1 shows an example of a binary image generated by the sensor and the related output data organization.

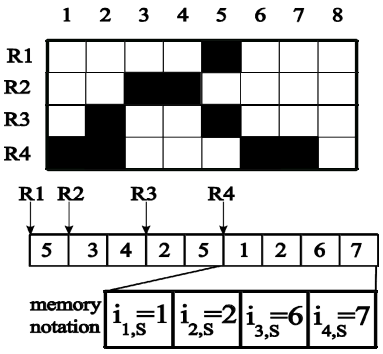


Figure 1. Example of contast-based binary image genarated by the vision sensor together with the related sensor output data, based on positional coding. The memory notation is used for the algorithm formalization.

Considering one row of the array, for each pixel the address is delivered and stored in the memory according with the raster-scan mode. If we consider the pixels within row R3 and R4, although pixels 2 and 5 occupy contiguous locations in the memory, they are physically placed at a distance of 3 pixels from each other in the array. One of the main issues here is that binary contrast is in fact a fairly poor information, which is not reliable enough for certain applications. In order to enrich the information of the system, a stereo approach has been investigated aimed at being applied in people counting and people detection applications. Due to the non-standard vision sensor adopted, a custom stereo algorithm needs to be developed, suitable to

978-1-4244-9137-7/11/$26.00 ©2011 IEEE

efficiently process data in the same order as they are dispatched by the sensor. In this paper, we want to demonstrate the feasibility of a stereo vision system based on an ultra-low power, low-resolution vision sensor, featuring binary contrast images. This objective can also be summarized as follows:

- low-resolution, custom vision sensor delivering minimum information about the scene (i.e. binary contrast) can still be used to retrieve depth information from the scene itself, with relatively low accuracy;
- address-driven data coding can be efficiently processed by custom stereo-vision algorithms, with low complexity and low energy-budget, suitable to be embedded into a standard processing architecture.

Most recently published results on a stereo-vision system, based on a custom address-driven vision chip, report an average power consumption of the order of 5W [5,6], which is in-line with other standard vision systems, but cannot be properly considered a low power system. The paper is organized as follows. Section II describes the system together with some geometry considerations. Section III describes the implemented stereo algorithm, while hardware issues are faced in Section IV. Experimental results and system performance are reported in Section V.

II. System Geometry

People Counting System

Focusing on a people counting application, some assumptions have to be made in order to reduce the overall problem complexity. Fig. 2 shows the people counting demonstrator, where the vision sensor has been installed overhead at about 3m height, monitoring an effective area of 2m x 1m referred to the floor. In this way, two persons have enough space to walk through at the same time. In case of a single sensor, due to the ambiguities coming from the binary information and the low imager resolution, it would be almost impossible to discriminate two or more persons walking close to each other. Moreover, shadows complicates things even further, as shown in the example of Fig. 3, where the person cannot be distinguished from the shadow. Using stereo, we aim at being able to distinguish the head of a person from its body by only selecting the information associated to a certain depth. Thus, we will also be able to get rid of the shadows. For this application, the required sensor depth resolution is not that large, ranging around 5 to 10 cm. These system specifications can therefore meet our sensor constraints.

Stereo-Vision Sensor

The two sensors are mounted on a PCB 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. It is worth noticing that 1.5cm at the floor is mapped into 1 pixel. In this phase of the work, many simplifications have been made, assuming the non-idealities of the optics (such as distortions and aberrations) to be negligible.

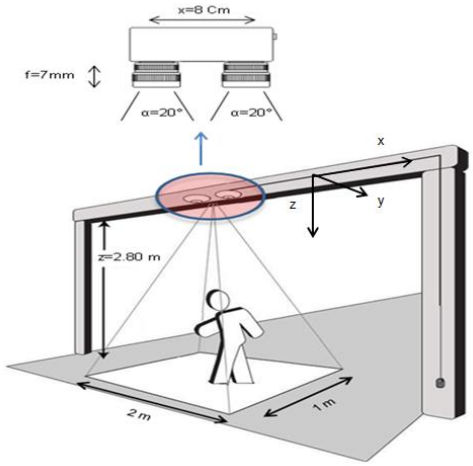


Figure 2. Prototype of the overhead people counting system. In the blow-up, the stereovision system is depicted with geometric details.

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 Fig. 2. We can determine the relationship between the optical system and the resulting stereo-vision system:

$$z = \frac{fb}{nP} \qquad (1)$$

where f is the focal length, b the baseline, P the pixel pitch and n the disparity expressed in pixels.

III. Disparity Algorithm

The proposed stereo-vision algorithm takes advantage from the characteristics of the sensor, which delivers data in a sparse-matrix through a positional coding. This makes the standard stereo area based [7,8,9,10] processing to be inefficient and extremely power hungry, due to the computation overhead required to re-organize data into a standard matrix. The novel stereo matching algorithm is based on the graph theory [11,12,13].

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.

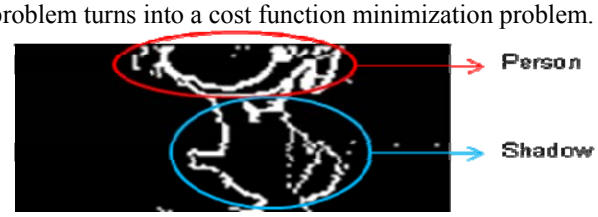


Figure 3. Example of a walking person with shadow.

For the sake of simplicity, let consider only the generic row (R) of the image NxM. Under the assumption that the two

sensors are properly calibrated on x-axis, as stated in Section II, a pair of corresponding points can be searched within the same row of the two sensors (left R_L and right R_R):

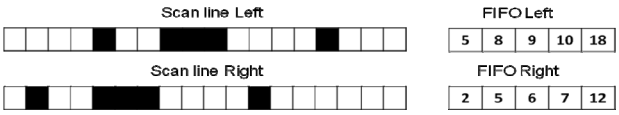


Figure 4. Example of ideal physical position in the images (Scan line) and the relative memory location in a FIFO.

Referring to FIFO Left, reported in a Fig.4, the pixel organization is therefore: ($i_{1,l}$=5,$i_{2,l}$=8 $i_{3,l}$=9,$i_{4,l}$=10, $i_{5,l}$=18). In this case, the best matching provided by the algorithm is:

$$S_{5,5} = (i_{1,l}, i_{1,r}), (i_{2,l}, i_{2,r}),..., (i_{5,l}, i_{5,r}) \quad (2)$$

$$= (5_l,2_r)\ (8_l,5_r)\ (9_l,6_r)\ (10_l,7_r)\ (18_l,12_r)$$

It is worth noticing, that in general the best matching is not necessary given by the pixels stored in the same left and right FIFO locations: ($i_{k+1,l}$, $i_{p+1,r}$)with $k \neq p$.The algorithm looks for the sequence of the best matching pixel pairs:

$$S_{j,p} = (i_{k,l}, i_{q,r})...(i_{j,l}, i_{p,r}) \quad \text{for } k,j,p,q \in \{1,2,\ldots,M\} \quad (3)$$

where $i_{j,l} > i_{k,l}$, $i_{p,r} > i_{q,r}$ and for any $k \neq j$ and $q \neq p$.

The best matching series is based on the following criteria which are embedded into a cost function:

I. Penalizing successive active pixels, which are physically far from each other on the same row of the sensors array .

II. Penalizing large disparities between pixels on left and right rows respectively: $d_{j,p} = l_{i_{j,l}} - r_{i_{p,l}}$.

The general cost function is:

$$C(S_{j,p}) = \sum_{1}^{M-1} {}_{j,p} \underbrace{A((i_{j,l} - i_{j-1,l} + 1) + (i_{p,r} - i_{p-1,r} + 1))}_{\mathrm{I}} + \underbrace{f(d_{j-1,p-1}, d_{j,p})}_{\mathrm{II}} \quad (4)$$

where the sequence S is expressed in eq.(3), the first term on the left implements the criterion I, the second term implements the criterion II. 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:

$$f(d_{j-1,p-1}, d_{j,p}) = \frac{|d_j - d_{j-1,p-1}|}{|l_{i_{j,l}} - l_{i_{j-1,p-1}}|} = \frac{|(l_{i_{j,l}} - r_{i_{p,l}}) - (l_{i_{j-1,l}} - r_{i_{p-1,l}})|}{|l_{i_{j,l}} - l_{i_{j-1,l}}|} \quad (5)$$

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:

$$A((i_{j+1,l} - i_{j,l} + 1) + (i_{p+1,r} - i_{p,r} + 1)) + f(d_{j,p}, d_{j+1,p+1}) \quad (6)$$

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.(6):

$$C(s_{j,p}) = C(s_{j-v,p-w}) + A((i_{j,l} - i_{j-v,l} + 1) + (i_{p,r} - i_{p-w,r} + 1)) + f(d_{j-v,p-w}, d_{j,p}) \quad (7)$$

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.(7) 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.(7) 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.(7).

Fig. 5 shows an example of cost function calculation under the following assumptions:

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

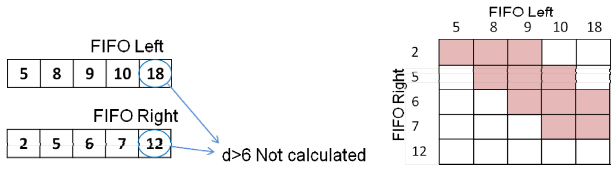


Figure 5. Cost function calculation

The cost function of the two rightmost pixels of the lines shown in Fig.5 (left) is not calculated, being their disparity larger than 6 pixels. The algorithm was tested using real date delivered by the vision sensor placed at the ceiling of a corridor and monitoring people walking through. As shown in Fig. 8 a) and b), the maximum number of detected edges in one row is less than 6. Fig. 6 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 M^4 with respect to the active pixels, while C2 is $O(M^2)$ to power two.

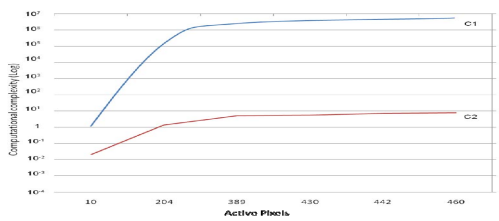


Figure 6. Computational compelxity

IV. SYSTEM ARCHITECURE

Figure 7a shows a simplified block-diagram of the stereo-vision system. It consists of the two vision sensors (SL, SR), connected with an FPGA, which provides the proper timing control forces both sensors to work synchronously, acquires images, reads binary data from both and sends them to a PC through a USB link. This is accomplished at a frame rate up to 25 fps. An FPGA-based developing board (XEM3001) was been used, mounting a medium-density Spartan-3 Xilinx component. Two 4Kx8bit FIFO memories blocks are synthesized in the FPGA. Here, data are packed row by row, separated by the End Of Row (EOR). Moreover, each new frame is separated by the End Of Frame (EOF). System control parameters is only the integration time. A simple user friendly interface has been realized with Labview. Binary frames are collected by the pc on which the C code algorithm runs in real time.

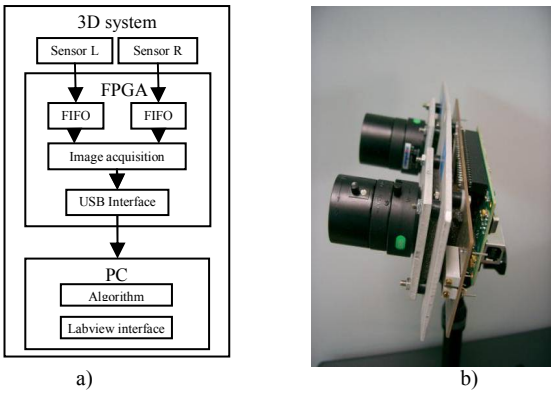
V. EXPERIMENTAL RESULTS

The system was installed in the corridor of our research center and benchmarks have been acquired related to the flux of people. Fig. 8 shows an example. Figure 8c) represents the disparity map obtained from 8a) and 8b). The algorithm, based on the graph theory, is able to find the best match, pixel by pixel, between the two images. For every pixel the disparity map (Fig. 7c) shows the disparity value, providing the height between the system and object. In Fig. 7c) different disparity levels are mapped in different colors. The resolutions, is regulated by:

$$\left|\frac{\partial z}{\partial \Delta v}\right| = \frac{d_{12} f}{p(\Delta v)^2} = \frac{z}{\Delta v} = \frac{z^2 p}{d_{12} f^2} \qquad (8)$$

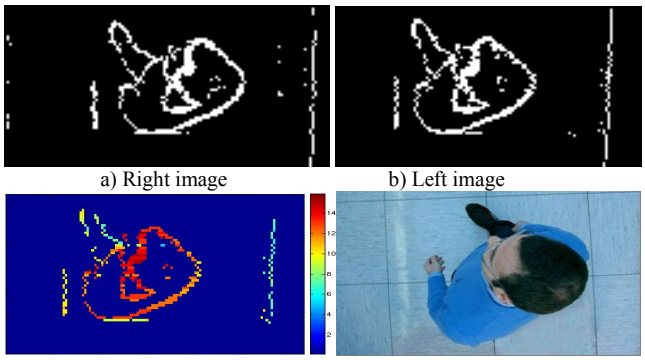
The system was installed at a height of around 3m and performs a resolution along z of about 5 cm near the head and about 40cm near the feet.

It is possible to adapt the resolution in z, changing the focal length and the height of the system.



a) b)

Figure 7. a) Block diagram of the stereo vision system; b) prototype.



a) Right image b) Left image

c) Disparity map d) Real scene (webcam)

Figure 8. Example of a snapshot processed with the proposed algorithm.

VI. CONCLUSION

The presented stereo vision adopts a novel algorithm for disparity estimation, based on the graph theory, where the minimum cost for any matching is computed through dynamic programming. A first prototype has been developed based on custom vision sensor with binary contrast extraction. The algorithm has been preliminarily tested in a overhead people counting application. Although the low pixel count of the binary sensor, the work demonstrates surprising performance reaching a depth resolution of 5cm over a measurement range from 1m to 3m. With proper assumptions, the developed algorithm can be implemented into a medium density low-cost FPGA.

VII. REFERENCES

[1] M. Gottardi, N. Massari, S.A. Jawed, "A 100uW 128x64 Pixels Contrast-Based Asynchronous Binary Sensor for Sensor Network Applications," IEEE Journal of Solid State Circuits, vol. 44, no. 5, pp. 1582-1592, May 2009.

[2] S. Hanson, D. Sylvester, "A 0.45-0.7V Sub-Microwatt CMOS Image Sensor for Ultra-Low Power Applications," Symp. on VLSI Circuits, 2009, pp. 176-177, 2009.

[3] K. Cho, D. Lee, J. Lee, G. Han,"Sub-1-V CMOS image sensor using time-based readout circuit," IEEE Transactions on Electron Devices, vol. 57, No.1, Jan. 2010, pp. 222-227.

[4] F. Tang, Y. Chao, A. Bermak, "An Ultra-Low Power Current-Mode CMOS Image Sensor with Energy Harvesting Capability," Proc. of European Solid State Circuits Conf., pp. 126-129, Sept. 2010.

[5] http://www.ait.ac.at/ SMART EYE Stereo Sensor

[6] Schraml, S.; Belbachir, A.N.; Milosevic, N.; Scho□n, P. , "Dynamic Stereo Vision System for Real-time Tracking" , Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on

[7] http://vision.middlebury.edu/stereo/

[8] C. L. Zitnick and T. Kanade "A Cooperative Algorithm for Stereo Matching and Occlusion Detection" Proc. IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 7, July 2000

[9] A. Fusiello, V. Roberto and E. Trucco, "Experiments with a new area-based stereo algorithm" Image Analysis and Processing, Lecture Notes in Computer Science, 1997, Volume 1310/1997, 669-676.

[10] T. Kanade and M. Okutomi, "A stereo matching algorithm with an adaptive window" IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 16, no. 9, pp. 920-932, 1994.

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

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

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