Proc. of the 6th WSEAS Int. Conf. on Signal Processing, Computational Geometry & Artificial Vision, Elounda, Greece, August 21-23, 2006 (pp192-197)

A low level real-time vision system using specific computing architectures

EDUARDO ROS, JAVIER DIAZ, SUHAIL M.I. ODEH, ANTONIO CAÑAS
Department of Computer Architecture and Technology,
University of Granada,
E.T.S.I.I.T., C/ Periodista Daniel Saucedo s/n, E-18071
SPAIN

Abstract: - In this work we present a vision system that includes circuits to compute different modalities in parallel, such as local image features (magnitude, phase and orientation), motion and stereo vision. This becomes possible by efficiently using the massive parallel computing resources of FPGA devices. The paper briefly describes the complete system and discusses the hardware consumption and performance of each visual modality. Finally, the work highlights that huge amount of data produced by such a system and the necessity of on-chip integration mechanisms.

Key-Words: - Real-time vision, motion, stereo, local image structure, FPGA

1 Introduction

After many years of research in the field of computer vision we are still far away from achieving outstanding vision skills similar to biological systems. Nevertheless, we have developed models for extracting robustly visual modalities such as stereo, motion, local features, etc; that are of high potential interest if we are able to use them in real-world applications. But most of these models require high computational load and cannot be processed in real-time using conventional computing platforms (single processor computers). This strongly limits their usability only to applications in which vision processing can be done off-line, while most of the real-world applications (vigilance, navigation, automatic object recognition, etc) require on-line vision processing.

This has motivated the design of specific computing architectures by different authors [1-10]. The implementation of real-time vision systems addresses different target fields and objectives:

a. *Real-time processing for embodied vision experiments.* Nowadays, it has become clear that simple off-line simulations are not enough to understand the way that different tasks are concurrently performed in the visual cortex. Furthermore, there is a strong working hypothesis called "embodiment concept" which states that any realistic simulation of a biologically inspired processing system should be tested in the framework of a certain task. The way that this task is achieved can be used to validate the different parts in which is based the success of the system. The embodiment concept is based on the hypothesis that biology has developed the impressively smart systems in nature through evolution trying to optimize certain tasks that improve the individual survival and specie perpetuation.

b. *Active vision.* The perception process is active. It combines sensori-motor capabilities in an integrative manner. Not only haptics but also vision is an active process in which intentional primitives drive certain mechanisms (such as fixation, smooth pursuing for stabilization, etc.) that enhance the accuracy of the system. Furthermore, it is also believed that attention is a useful mechanism in order to achieve very high performance with constrained processing resources. But active perception processes can only be studied in the framework a perception-action closed-loops. This specifically requires real-time processing and represents a strong motivation for developing high performance vision processing architectures.

c. *Understanding by building.* From an engineering point of view, we only fully understand certain mechanisms if we are able to implement them. In the framework of computer vision systems, as engineers, trying to build efficient image processing architectures based on biological vision systems is a very interesting approach since we face the same limitations as nature also with constrained processing resources. According to the "neuromorphic

engineering" paradigm we adopt an opportunistic attitude in which we try to emulate schemes that seem to be efficient in the biological systems and we avoid other features that are more intrinsically related with the tissues in which they are based. Furthermore, not being limited by some biological restrictions (such as power or conductance and switching capability of neuron wiring and connections) we can take full advantage of certain outstanding characteristics of electrical technology, such as high communication bandwidth, high speed state switching, etc.

d. *Smart vision systems for real world applications.* Real-time processing of local features, motion and stereo is interesting for a wide range of applications in real world scenarios. Therefore, the implementation of high-performance computing architectures has an interest in itself for solving real world problems.

In this paper we present a system composed from deep pipelined datapaths for the computation of local features, stereo and motion. The implementation of computing architectures that efficiently take advantage of the large amounts of parallel computing resources of FPGA devices for vision schemes is not common. It requires a very well structured design strategy in order to arrive at datapaths delivering one image fearture estimation (for example disparity or local velocity) per clock cycle. Current FPGA devices include several millions of configurable gates that can be used to implement very efficient computing architectures.

An efficient implementation of an algorithm in specific hardware requires modifying the original model (at least translating most of the computations to a custom arithmetic). This is done by bit cutting strategies that allow optimizing the hardware resources but also affect the accuracy of the system. Because of this, the final implemented circuit can be seen as a new model that needs to be evaluated with benchmark sequences to test the accuracy vs. efficiency trade-off of the system.

The main purpose of the paper is to illustrate how different visual modalities can be processed in real-time on the same chip using the currently available parallel resources of FPGA devices. Nevertheless, we want to highlight that despite the significant increment in computing power of these chips, taking full advantage of different visual modalities available on the same device is difficult since transmission bandwidth constraints limit the amount of data that can be transferred between different computing platforms (or chips). This technology limitation makes interesting to study efficient integration mechanisms that can be implemented on the same chip leading to sparse multimodal entities encoding the maximum information for higher visual stages towards scene understanding. The efforts to build a real-time vision machine using specific hardware require integration mechanisms leading to meaningful multimodal entities that can be efficiently transferred to other computing chips. These integration mechanisms need to be hardware friendly because they need to be included within the same chip in which the early visual primitives are extracted.

Furthermore, the necessity of "scene understanding" into the chips fits well with the technology trends. Current reconfigurable devices include (hard and soft) processors which are a more suitable hardware choice for extracting the meaningful data computed at the early vision stages. As it occurs in the brain, the system requires high parallelism in early stages where primary rear data is processed and less computing speed is required for more sequential tasks involved in other layers dealing with information of higher abstraction levels.

2 Systems description

The presented system includes different visual modalities (as illustrated on Fig. 1). In the results section we indicate the hardware resources consumed by each visual modality and the necessity of the integration module to reduce the output transference bandwidth. Fig.2 shows the parallel datapaths designed for each visual modality.

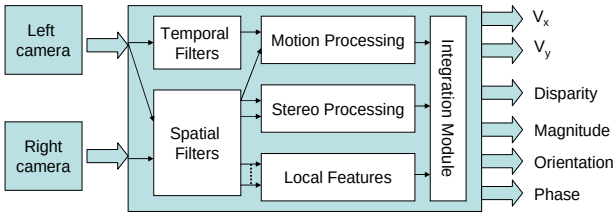


Fig. 1. Full system on chip. It only requires external memory resources (not shown in the Figure) for temporal variables.

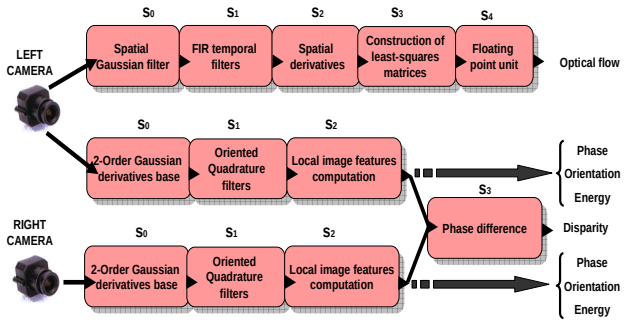
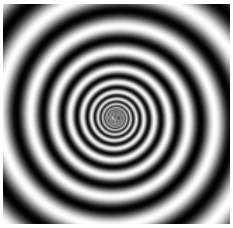


Fig. 2. Parallel pipelined datapaths for each visual modality.

2.1 Local features

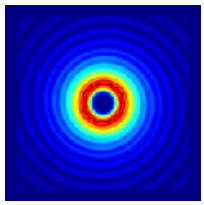
We have implemented a system to extract the local structure of the images [3]. We have used steerable filters (based in second order Gaussian derivatives) to compute the magnitude, phase and orientation of each pixel.

(3.a)



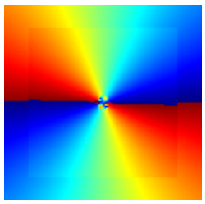
Magnitude

(3.b)



Orientation

(3.c)



Phase

(3.d)

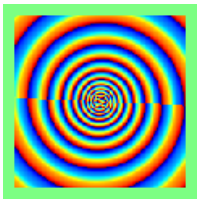


Fig.3. Local features extracted with the hardware system from a synthetic spiral image.

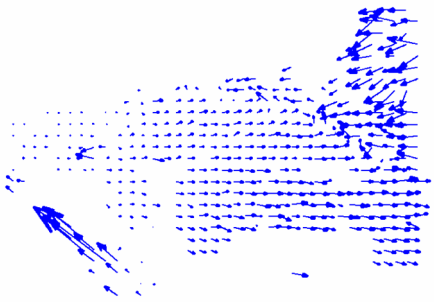
See Fig. 3 for some results (only hardware results are shown to illustrate the output data streams of the system). We use the results computed out of a synthetic image to facilitate their visual evaluation. Fig. 3 shows how the results are accurate despite the restricted computation precision.

2.2 Motion processing system

We have designed a superpipelined motion processing system with an outstanding computing power. We have implemented in specific hardware the Lucas & Kanade algorithm [11] with the modifications suggested by Brandt [12]. Fig. 4 shows some qualitative results. Note that there is no significant degradation in the hardware results due to the restricted computation precision of the circuits that use fixed point arithmetic. Details about the hardware implementation of the optic flow system can be found in [1].

(4.a)

(4.b)



(4.c)

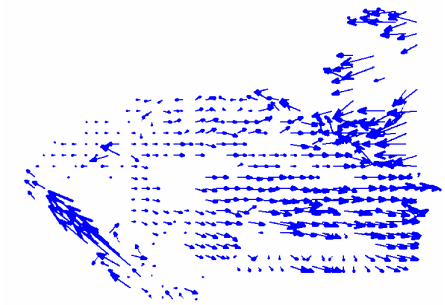


Fig. 4. Qualitative optic flow results. Overtaking sequence from the rear view mirror. a) Original sequence frame. b) Software results. c) Hardware results.

2.3 Stereo Processing system

We have implemented a superpipelined datapath that efficiently computes a hardware friendly phase-based stereo algorithm [13]. Details about the hardware implementation of the stereo system can be found in [2]. Fig. 5 shows some qualitative results. Note that the only difference between the software and hardware results is an increase in the salt and pepper noise in areas without structure due to the restriction in precision of the hardware-based computations. This artefact is easy to filter using structure-based confidence measures.

(5.a)

(5.b)

(5.c)

Fig. 5. Stereo processing qualitative results. a) Original image. b) Software results. c) Hardware results. Depth is encoded in grey levels.

3 Results

Different modalities can be processed in parallel on the same chip using similar spatio-temporal convolution kernels.

We focus on the following visual modalities:

- Local features (magnitude, phase and orientation).
- Motion.
- Stereo.

Fig. 6 shows the hardware resources consumed by each visual modality (note that all of them fit into the same chip). It shows the amount resources of different kinds: general purpose logic (slices), embedded memory (EMB) and embedded multipliers (E_Mult). Intputs circuits include the video frame-grabber, VGA output for visualization, memory management units as well as user interface for parameter adaption.

Table 1 includes results about the performance in terms of computing speed (Kpps stands for Kpixels per second). We have constrained all the circuits to process at 45500 Kpps (the datathroughput of limiting circuit which is the motion core). Since the three cores are fed by an on-chip frame-grabber that can be particularized for each modality. It may be desirable to process motion at more frames per second (than the standard 30 fps of conventional cameras) if we use specific oversampled sensors. On the other hand, stereo and the local features may benefit of a higher spatial resolution (temporal aliasing does not affect these pure spatial modalities).

Table 1. System performance and output data bandwidth with an input image resolution of 1000x1000 pixels. (* indicates that we also include 8 bits of bitwidth of the input image that are also transferred in the output data bandwidth).

	Computing speed (Kpps)	**Output bitwidths (bits)**	**Output data bandwidth (MB/s)**
Motion	45500	$12(V_x)+12(V_x)$	133.3
Stereo	45500	12 (Disparity)	95.2
Local features	45500	9(M)+9(P)+8(O)	179.3
Full system		70*	388.8*

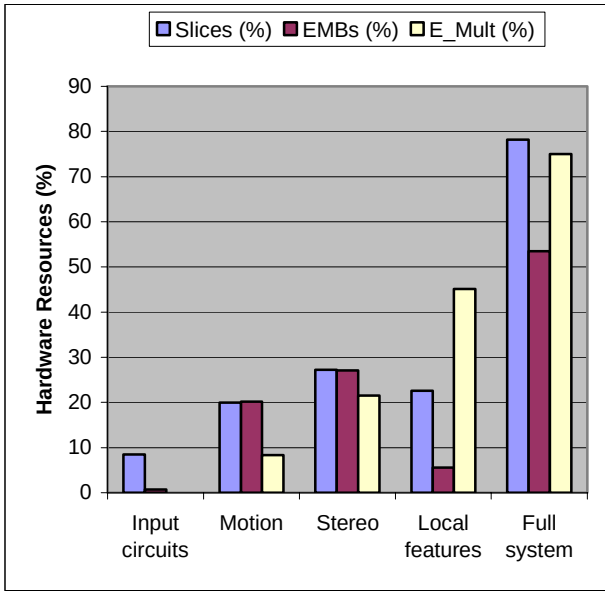


Fig. 6. Hardware resources consumption. We indicate the different types of hardware resources of a Virtex II XC2V6000-4 [14] consumed by each visual modality. All of them fit together into a single chip.

4 Conclusion

This paper has illustrated how different visual modalities can now be efficiently computed on a single chip. This is of high interest for certain applications in which a specific vision modality can be crucial for solving a real-world application (for instance motion for car overtaking scenarios [15, 16]). But on the other hand, although with similar primitives (spatio-temporal convolution kernels) can extract different vision modalities on the same chip they are of little interest if communication bandwidth constraints limit their transference to other computing platforms. Even if they could be transferred the receiving computing platform would be overloaded just retrieving such a large amount of data. Therefore, in this framework hardware friendly integration schemes that allow clustering all this information into sparse multimodal entities becomes of extreme interest. In fact, this kind of schemes should be embedded into the same chip in which the visual primitives are extracted.

The main conclusion of this paper is that although specific hardware is of high interest to efficiently extract early and dense visual primitives a multimodal vision system represents such a diverging data structure in which on-chip convergence (multimodal integration mechanisms) becomes necessary. There are already compacting schemes that use integration mechanisms to cluster the different visual modalities into specific multimodal entities [17]. In the close future work we will investigate how to implement these schemes also in the same chip to overcome the inter-chip bandwidth limitations. It is important to note that although high performance I/O resources (for instance Rocket I/O in Xilinx devices) provide very high bandwidth if the data transferred needs to be further processed on other platforms it will easily overload the receiving system. In our case we are working with PCI platforms, with advanced PCI express ports (several channels) we would be able to carry all the output data stream but the PC receiving the data would get overloaded and therefore further processing in real time is not possible unless we include compacting schemes on-chip.

This problem also leads to stand-alone systems where the problem and its solution are addressed into the same device, for example in the framework of intelligent vehicles [15, 16]. In the case of many stand-alone systems they only deliver specific alarm signals when detecting concrete situations. In this way the bandwidth is significantly reduced, using only output basic command and data transmission.

While we talk about early cognitive vision, middle vision and high level cognitive vision layers to refer to the processed information of different abstraction level. Technology limitations make a strong difference, not only in the abstraction level of the information that is being treated but also on the amount of data that each layer handles. In this paper, it has become clear that while early cognitive vision modalities should be as dense (stereo, motion, local features) as the local structure allows higher visual layers need to efficiently handle a large amount of data, therefore integration mechanisms become of extreme interest. This should also hold true for biological systems in which as we go to higher abstraction vision layers information becomes more concrete and sparse. While early cognitive vision deals with pixels as input stream higher layers shall deal with multimodal vision entities.

Acknowledgements

This work has been supported by the Spanish Grant (DPI2004-07032) and the EU grant DRIVSCO (IST-016276-2).

References

[1] J. Díaz, E. Ros, F. Pelayo, E. M. Ortigosa and S. Mota, "FPGA based real-time optical-flow system," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16,

no. 2, pp. 274-279, 2006.

[2] J. Díaz, E. Ros, R. Carrillo and A. Prieto, "Real-time system for high-image-resolution disparity," *IEEE Trans. On Image Processing*, 2006, accepted for publication.

[3] J. Díaz, "Multimodal bio-inspired vision system. High Performance motion and stereo processing architecture", PhD Thesis dissertation. University of Granada. 2006.

[4] J.L. Martín, A. Zuloaga, C. Cuadrado, J. Lázaro, and U. Bidarte, "Hardware implementation of optical flow constraint equation using FPGAs," *Computer Vision and Image Understanding*, vol.98, no. 3, pp. 462-490, June 2005.

[5] M. V. Correia, and A. C. Campilho, "Real-time implementation of an optical flow algorithm," in *Proc. international Conference on Pattern Recognition (ICPR2002)*, 2002, p. 40247, vol. 4. pp. 247-250.

[6] P. Cobos, and F. Monasterio. "FPGA implementation of the Horn & Shunk Optical Flow Algorithm for Motion Detection in real time Images," in *Proc. of the XIII Design of Circuits and Integrated Systems Conference*, Madrid, Spain. Nov. 1998, pp. 616-621.

[7] H. Niitsuma, and T. Maruyama, "Real-Time Detection of Moving Objects," *Lecture Notes in Computer Science, FPL 2004*, vol. 3203, pp. 1153-1157, September 2004.

[8] A. Darabiha, W. J. MacLean and Jonathan Rose, "Reconfigurable Hardware Implementation of a Phase-Correlation Stereo Algorithm," *Machine Vision and Applications Journal*, March, 2006.

[9] J. Woodfill, B. Von Herzen, "Real-time stereo vision on the PARTS reconfigurable computer", presented at the Conf. IEEE Symposium on Field-Programmable Custom Computing Machines, Napa, April 1997, pp. 201.

[10] T. Kanade, "Development of a Video-Rate Stereo Machine," in *Proc. of the 1994 ARPA Image Understanding Workshop* (IUW'94), November, 1994, Monttey Ca, pp. 549–558.

[11] B.D. Lucas, and T. Kanade: "An Iterative Image Registration Technique with an Application to Stereo Vision," In *Proc. of the DARPA Image Understanding Workshop*, April 1981, pp. 121-130.

[12] J.W. Brandt, "Improved Accuracy in Gradient Based Optical Flow Estimation," *International Journal of Computer Vision*, vol. 25, issue 1, pp. 5-22, October 1997.

[13] F. Solari, S. P. Sabatini, G. M. Bisio, "Fast technique for phase-based disparity estimation with no explicit calculation of phase," *Electronics Letters*, vol. 37, issue 23, pp. 1382 -1383, 2001.

[14] Xilinx Virtex II FPGAs, [Online]. Available: http://www.xilinx.com/products/silicon_solutions/fpgas/virtex/virtex_ii_platform_fpgas/resources/index.htm.

[15] S. Mota, E. Ros, J. Díaz, E. M. Ortigosa, and A. Prieto, "Motion-Driven Segmentation by Competitive Neural Processing," *Neural Processing Letters*, vol.22, no 2, pp. 125-147, 2005.

[16] J. Díaz, E. Ros, S. Mota and R. Agis, "Real-time embedded system for rear-view mirror overtaking car monitoring," *Lecture Notes On Computer Science*, Springer-Verlag, 2006, to be published.

[17] N. Krüger and M. Felsberg, "An Explicit and Compact Coding of Geometric and Structural Information Applied to Stereo Matching," *Pattern Recognition Letters*, vol. 25, Issue 8, pp. 849-863, June 2004.