FPGAs and other reconfigurable architectures are becoming prolific as a platform for implementing a broad domain of applications. In this session, we have three papers and an interactive presentation focused on the design of computer vision, machine learning, and video processing on reconfigurable architectures.

**11.6.1 Efficient FPGA Acceleration of Convolutional Neural Networks Using Logical-3D Compute Array**

**Speaker:** Atul Rahman, UNIST, KR  
**Authors:** Atul Rahman1, Jongeun Lee1 and Kiyoung Choi2  
1UNIST, KR; 2Seoul National University, KR  
**Abstract**  
Convolutional Deep Neural Networks (DNNs) are reported to show outstanding recognition performance in many image-related machine learning tasks. DNNs have a very high computational requirement, making accelerators a very attractive option. These DNNs have many convolutional layers with different parameters in terms of input/output/kernel sizes as well as input stride. Design constraints usually require a single design for all layers of a given DNN. Thus a key challenge is how to design a common architecture that can perform well for all convolutional layers of a DNN, which can be quite diverse and complex. In this paper we present a flexible yet highly efficient 3D neuron array architecture that is a natural fit for convolutional layers. We also present our technique to optimize its parameters including on-chip buffer sizes for a given set of resource constraint for modern FPGAs. Our experimental results targeting a Virtex-7 FPGA demonstrate that our proposed technique can generate DNN accelerators that can outperform the state-of-the-art solutions, by 22% for 32-bit floating-point MAC implementations, and are far more scalable in terms of compute resources and DNN size.

**11.6.2 Energy Efficient Video Fusion with Heterogeneous CPU-FPGA Devices**

**Speaker:** Peng Sun, University of Bristol, GB  
**Authors:** Peng Sun1, Alin Achim1, Ian Hasler, Paul Hill2 and Jose Nunez-Yanez1  
1University of Bristol, GB; 2Qioptiq LTD, GB  
**Abstract**  
This paper presents a complete video fusion system with hardware acceleration and investigates the energy trade-offs between computing in the CPU or the FPGA device. The video fusion application is based on the Dual-Tree Complex Wavelet Transforms (DT-CWT). Video fusion combines information from different spectral bands into a single representation and advanced algorithms based on wavelet transforms are compute and energy intensive. In this work the transforms are mapped to a hardware accelerator using high-level synthesis tools for the FPGA and also vectorized code for the single instruction multiple data (SIMD) engine available in the CPU. The accelerated system reduces computation time and energy by a factor of 2. Moreover, the results show a key finding that the FPGA is not always the best choice for acceleration, and the SIMD engine should be selected when the wavelet decomposition reduces the frame size below a certain threshold. This dependency on workload size means that an adaptive system that intelligently selects between the SIMD engine and the FPGA achieves the most energy and performance efficiency point.

**11.6.3 Highly Efficient Reconfigurable Parallel Graph Cuts for Embedded Vision**

**Speaker:** Antonis Nikitas, Technical University of Crete, GR  
**Authors:** Antonis Nikitas1 and Ioannis Papaefstathiou2  
1Technical University of Crete, GR; 2Synelixis Solutions Ltd, GR  
**Abstract**  
Graph cuts are very popular methods for combinatorial optimization mainly utilized, while also being the most computational intensive part, in several vision schemes such as image segmentation and stereo correspondence; their advantage is that they are very efficient as they provide guarantees about the optimality of the reported solution. Moreover, when those vision schemes are executed in mobile devices there is a strong need, not only for real-time processing, but also for low power/energy consumption. In this paper, we present a novel architecture for the implementation, in reconfigurable hardware, of one of the most widely used graph cuts algorithms, which is also the fastest sequential one, called BK. Our novelty comes from the fact that we use a 2-level hierarchical decomposition method to parallelize it in a very modular way allowing it to be efficiently implemented in FPGAs with different number of logic cells and/or memory resources. We fast-prototyped the architecture, using a High level synthesis workflow, in a state-of-the-art FPGA device; our implementation outperforms an optimized reference software solution by more than 6x, while consuming 35 times less energy. To the best of our knowledge this is the first parallel implementation of this very widely used algorithm in reconfigurable hardware.
A NOVEL BACKGROUND SUBTRACTION SCHEME FOR IN-CAMERA ACCELERATION IN THERMAL IMAGERY

Speaker: Konstantinos Makantasis, Institute of Communication and Computer Systems, GR

Authors: Antonis Nikitakis¹, Ioannis Papaefstathiou², Konstantinos Makantasis³ and Anastasios Doulamis⁴
¹Technical University of Crete, GR; ²Synelixis Solutions Ltd, GR; ³Institute of Communication and Computer Systems, GR; ⁴National Technical University of Athens, GR

Abstract
Real-time segmentation of moving regions in image sequences is a very important task in numerous surveillance and monitoring applications. A common approach for such tasks is the "background subtraction" which tries to extract regions of interest from the image background for further processing or action; as a result its accuracy as well as its real-time performance is of great significance. In this work we utilize a novel scheme, designed and optimized for FPGA-based implementations, which models the intensities of each pixel as a mixture of Gaussian components; following a Bayesian approach, our method automatically estimates the number of Gaussian components as well as their parameters. Our novel system is based on an efficient and highly accurate on-line updating mechanism, which permits our system to be automatically adapted to dynamically changing operation conditions, while it avoids over/under fitting. We also present two reference implementations of our Background Subtraction Parallel System (BSPS) in Reconfigurable Hardware achieving both high performance as well as low power consumption; the presented FPGA-based systems significantly outperform a multi-core ARM and two multi-core low power Intel CPUs in terms of energy consumed per processed pixel as well as frames per second. Moreover, our low-cost, low-power devices allow for the implementation, for the first time, of a highly distributed surveillance system which will alleviate the main problems of the existing centralized approaches.

RADIATION-HARDENED DSP CONFIGURATIONS FOR IMPLEMENTING ARITHMETIC FUNCTIONS ON FPGA

Speaker: Felipe Serrano, Universidad Complutense de Madrid, ES

Authors: Marcos Sanchez-Elez, Inmaculada Pardines, Felipe Serrano and Hortensia Mecha, Universidad Complutense de Madrid, ES

Abstract
This paper presents a study of different implementations of arithmetic operations on FPGAs. Radiation vulnerability has been analyzed for each implementation using the fault injection platform NESSY. Results in terms of area, delay and reliability are presented. Taking into account the performed tests we propose to build a library of HDL templates. This library is used during the design process with a synthesis tool that implements digital circuits as reliable as possible. Experimental results show that those implementations using DSP slices are the ones which achieve better results.

CONFIGURATION PREFETCHING AND REUSE FOR PREEMPTIVE HARDWARE MULTITASKING ON PARTIALLY RECONFIGURABLE FPGAS

Speaker: Ann Gordon-Ross, University of Florida, US

Authors: Aurelio Morales-Villanueva, Rohit Kumar and Ann Gordon-Ross, University of Florida, US

Abstract
Partially reconfigurable (PR) FPGAs enable preemptive hardware (HW) multitasking using PR regions (PRRs). To enable this multitasking, the HW task's partial bitstream is downloaded to only the task's PRR, and only that PRR is reconfigured. Since only a small portion of the FPGA fabric is reconfigured, reconfiguration time is significantly reduced as compared to reconfiguring the entire fabric, however this time is not negligible. Reconfiguration time can be reduced/hidden using two techniques: configuration prefetching and configuration reuse. Even though these techniques can effectively reduce/hide reconfiguration overhead, prior works in preemptive HW multitasking did not use these techniques. To the best of our knowledge, no prior work evaluated physical implementations of these techniques on PR FPGAs, which precludes consideration of physical-implementation-specific details, such as delays in accessing bitstreams, speed limitations during reconfiguration, etc. In this work, we present a novel implementation of configuration prefetching and reuse for preemptive HW multitasking on a Virtex-5 FPGA, however, our established fundamentals are device-family independent.

Source URL: https://www.date-conference.com/date16/conference/session/11.6