The Collaborative Workshop on Architectures of Smart Cameras

WASC 2019

Presentations Abstracts

July 1-2, INSA Rennes

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES RENNES
IETR
VAADER
ACHIEVE ITN
Horizon 2020
European Union Funding
for Research & Innovation

On the Balanced Allocation of Convolutional Neural Network Models on FPGAs

A. Muñío-Gracia, J. Fernández-Berni, R. Carmona-Galán, and Á. Rodríguez-Vázquez
Instituto de Microeléctronica de Sevilla (IMSE-CNM), CSIC-Univ. Sevilla, Spain

Abstract: Deep Learning (DL) algorithms have demonstrated their competence in accurately extracting information from data, especially in the field of computer vision. DL has emerged as an end-to-end approach based on learned multi-level scene representations. A number of open-source frameworks have been created to describe convolutional neural network (CNN) models —a class of the deep neural networks (DNNs) that support DL. Their computational complexity prompts for hardware acceleration. The challenge in the design of hardware accelerators for CNNs is providing a sustained throughput with low power consumption. In order to test our architectural proposals, we will be employing FPGAs. They are reconfigurable, efficient, and have adjustable precision. FPGAs permit architectural exploration with shorter development time and lower cost than ASICs. This work introduces an scalable, framework-agnostic, architecture whose behavior self-adapts to the selected CNN configuration. A design space analysis is performed for some state-of-the-art CNNs, namely VGG-16, Tiny DarkNet, and SqueezeNet. The objective is a balanced allocation of resources. For this, tiling parameterization will be optimized attending to decisive performance criteria such as the number of memory accesses, data movement policy and throughput.

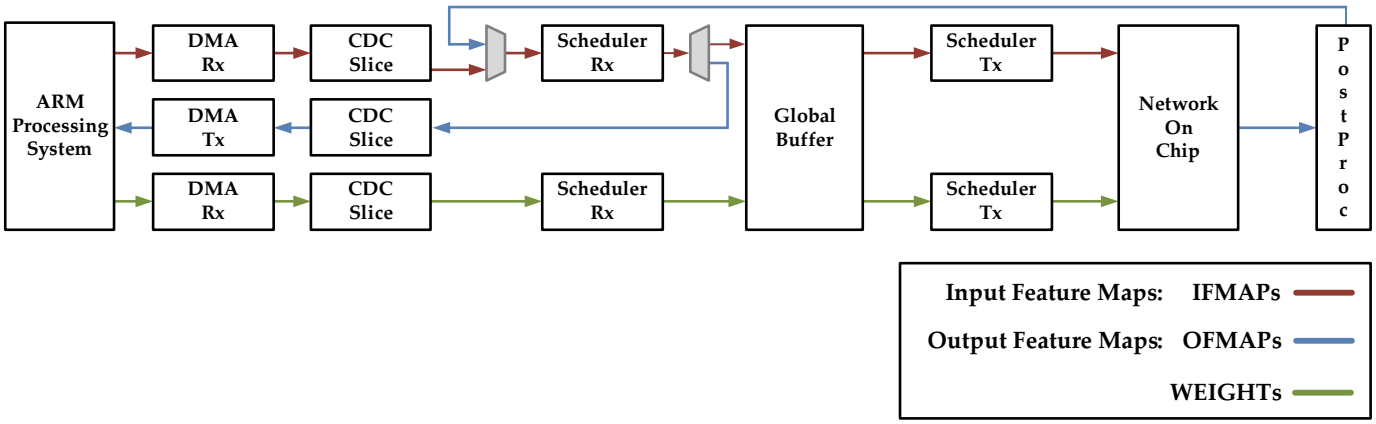


Fig. 1: General block diagram of FPGA Convolution Neural Network Accelerator

Acknowledgments: This work was supported by EU H2020 MSCA through Project ACHIEVE-ITN (Grant No 765866), by the Spanish MINECO and European Region Development Fund (ERDF/FEDER) through Project RTI2018-097088-B-C31 and by the US Office of Naval Research through Grant No. N00014-19-1-2156.

Towards a Simplified Procedure for CNN Performance Prediction on Embedded Platforms

D. Velasco-Montero, J. Fernández-Berni, R. Carmona-Galán, and Á. Rodríguez-Vázquez
Instituto de Microeléctronica de Sevilla (IMSE-CNM), CSIC-Univ. Sevilla, Spain

Abstract: Vision is arguably the technical field benefiting the most from the renaissance of artificial intelligence in the last few years. In particular, the convergence of massive datasets for training, boosted computational power, and enhanced machine learning techniques has given rise to highly accurate vision algorithms – even outperforming humans in certain tasks – based on convolutional neural networks (CNNs). The potential of these algorithms has attracted attention from many parties, both in academia and industry, spurring the development of a myriad of hardware platforms and software frameworks. The challenge now is how to efficiently leverage and integrate this variety of components in practical realizations, taking also into account that CNN models keep evolving at a rapid pace. With this scenario in mind, we have been working on a simplified procedure to predict the performance of CNNs running on embedded platforms in terms of throughput and power consumption. The objective is to facilitate the evaluation of the aforementioned components and CNN models prior to actually implementing them, thereby speeding up the deployment of optimal solutions. In this talk, we will describe key aspects of the proposed procedure. Specifically, we will elaborate on SweepNet, a deep neural network tailored for meaningful per-layer characterization. The performance models extracted from SweepNet for a hardware platform allow to accurately predict layer by layer the execution time and energy consumption of any other CNN running on that platform. As an example, Fig. 1 compares the time actually required to complete each layer of Network-in-Network (NiN) – a popular CNN model – on a Raspberry Pi 3B, with the time predicted from the characterization of SweepNet.

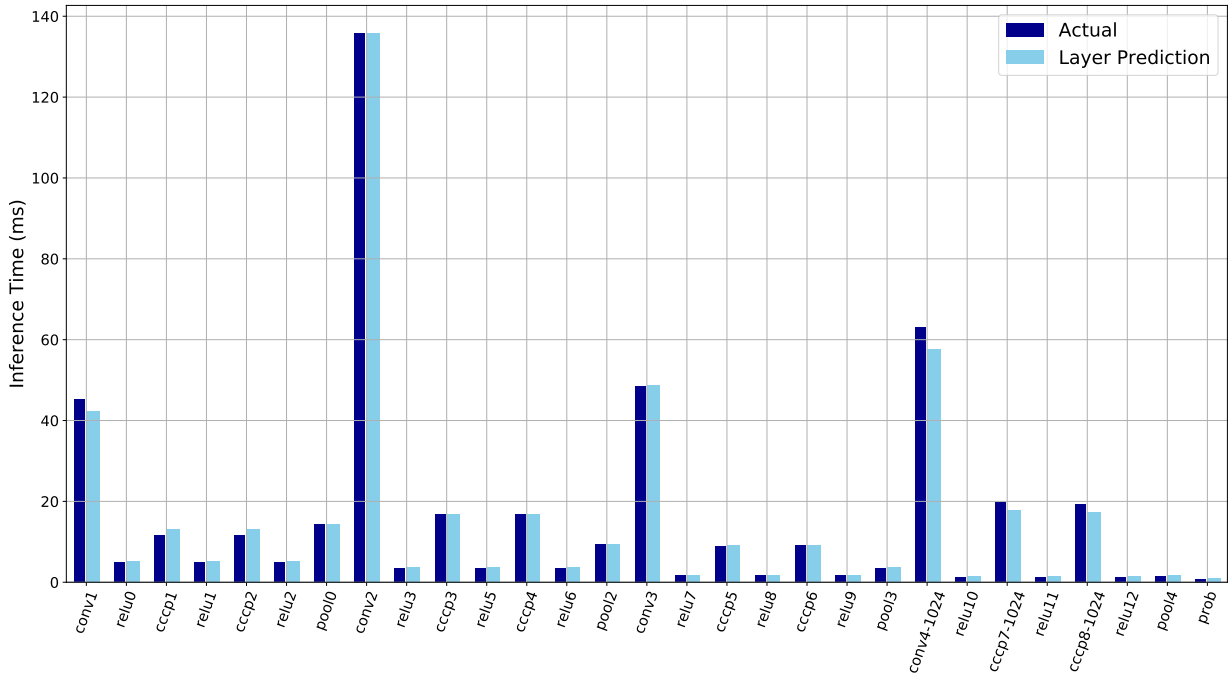


Fig. 1: Comparison of actual execution time of NiN layers vs. predicted time from SweepNet's characterization.

Acknowledgments: This work was supported by EU H2020 MSCA through Project ACHIEVE-ITN (Grant No 765866), by the Spanish MINECO and European Region Development Fund (ERDF/FEDER) through Project RTI2018-097088-B-C31, and by the US Office of Naval Research through Grant No. N00014-19-1-2156.

Towards Accurate Single-Stream Human Action Detection in Real-Time

Yu Liu, Fan Yang, Dominique Ginhac
Laboratoire ImViA, EA7535, Univ. Bourgogne Franche-Comté, Dijon, France

Abstract: Analyzing videos of human actions involves understanding the spatial and temporal context of the scenes. State-of-the-art action detection approaches have demonstrated impressive results using Convolutional Neural Networks (CNNs) within a two-stream framework. However, most of them operate in a non-real-time, offline fashion, thus are not well-equipped in many emerging real-world scenarios such as autonomous driving and public surveillance. In addition, they are computationally demanding to be deployed on devices with limited power resources (e.g., embedded systems). To address the above challenges, we propose an efficient single-stream action detection framework by exploiting temporal coherence between successive video frames. This allows CNN appearance features to be cheaply propagated by motions rather than being extracted from every frame. Furthermore, we utilize implicit motion representation to amplify appearance features. Our method based on motion-guided and motion-aware appearance features is evaluated on public dataset. Experiments indicate that the proposed method can achieve real-time action detection up to 32 fps with a comparable accuracy as the two-stream approach.

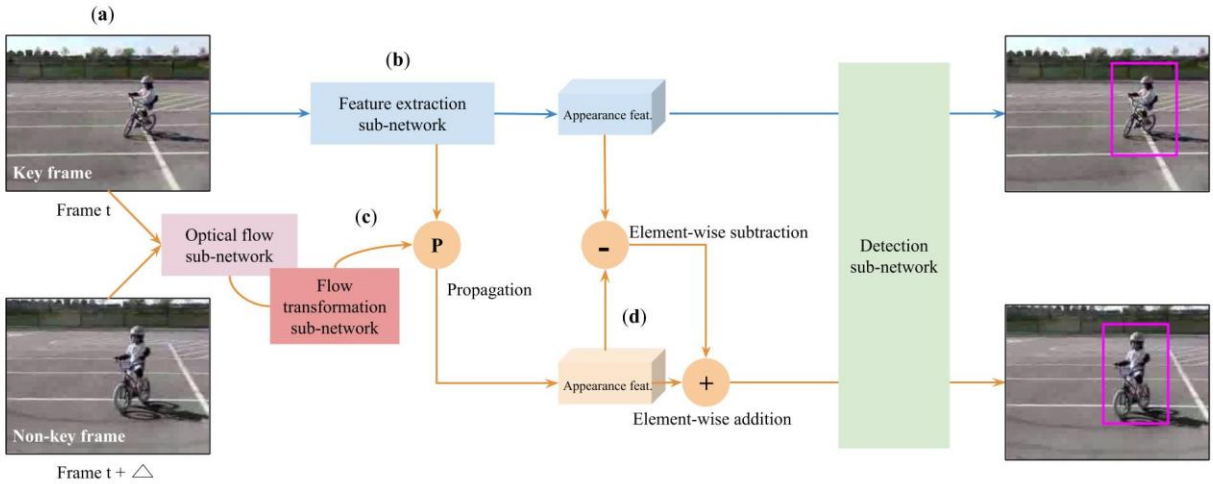


Fig. 1. Illustration of our motion-guided action detection framework with motion-aware appearance features.

Acknowledgments: This work was supported by the H2020 Innovative Training Network (ITN) project ACHIEVE (H2020-MSCA-ITN-2017: agreement no. 765866).

Methods for Autonomous Navigation and Localization in Traffic Environments

Gopi Krishna Erabati and Helder Araujo
Institute of Systems and Robotics (ISR), University of Coimbra, Portugal

Abstract: In the past few decades, systems particularly vehicles are focused to be automated to trim down the cause of accidents by humans. The consequences of the actions (for instance recklessness) by humans, may vary from damage to property to loss of life. Scientific and technical community has come up with many solutions (multi-domain) to these problems during the era of automation. Advanced driver-assistance systems (ADAS) and self-driving (autonomous) vehicles are interesting solutions to such problems. Every autonomous system have a classical sensing, processing and acting architecture. The role of computer vision community is to sense the environment with different sensors (specially optics) and process the information to understand the scene and thereby handing over significant information to the control expertise community to act on the system and control its behavior.

This project aims to play a role in computer vision expertise domain in the autonomous vehicles. Optical sensors (like 3D cameras) are used to acquire the data from the environment for inference. The processing of this data is limited to dynamic obstacle detection, visual odometry and partial SLAM in this project and the present focus of the work is on dynamic obstacle detection task. Robust and reliable algorithms with real time performance are vital for obstacle detection in traffic environments. With the advent of neural network community, many problems in the world are being solved by deep learning and obstacle detection is one among them. A way of taxonomy of detection algorithms is two stage and single stage methods. Two stage detection techniques have a region proposal stage and a classification stage like Faster-RCNN. Single stage detection techniques look through the input only once and detect the objects in the scene, like You Only Look Once (YOLO) and Single Shot Detector (SSD). The detection with Faster-RCNN is computed with VGG16 and ResNet50, YOLOv3 with ResNet50 and MobileNetv1 and SSD512 with VGG16, ResNet50 and MobileNetv1 as backbone networks to the main algorithm. Faster-RCNN is an improved version of R-CNN and Fast-RCNN, wherein, multiple convolution networks (convnet) are replaced by single convnet and selective search for region proposal is replaced by region proposal network (RPN) respectively. RPN accelerates the training and testing phases and improves performance. YOLOv3 directly predicts bounding boxes and class probabilities with a single network in single evaluation. The simplicity of this allows real time performance. Similar to YOLOv3, SSD detector detects bounding boxes and class probabilities with a end-to-end CNN architecture. There is always a tradeoff between accuracy and speed for algorithms. The two stage techniques work slightly better than single stage techniques but they are not comparable to real time processing. On the other hand, single shot techniques work nearly and above real time speeds and are comparable in accuracy with two stage techniques. The two stage and single stage techniques are trained with PASCAL VOC 2007 and 2012 training and validation sets and inference is done on PASCAL VOC 2007 test set on Google Colab K80 GPU. The networks to use 3D data, detect obstacles and understand the scene are being explored and architectures for the same will be developed in future.

Acknowledgements: This project is funded by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No. 765866.

A CMOS vision sensor for background subtraction on the focal plane

Daniel García-Lesta[1], Víctor M. Brea[1], Paula López[1], Diego Cabello[1], Stephen Carey[2], and Piotr Dudek[2]

[1] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain
[2]School of Electronic and Electrical Engineer, The University of Manchester, Manchester, United Kingdom

Abstract

Background subtraction is one of the first steps in different video processing algorithms. Thus, a real-time processing with low power consumption is convenient for different applications where power hungry devices with high computational capabilities can not be deployed. This work presents the design of a 24x56 pixel proof-of-concept CMOS vision sensor implementing the foreground detection algorithm Hardware Oriented Pixel Based Adaptive Segmenter (HOPBAS) on the focal plane, showing the operational blocks used and how they are implemented.

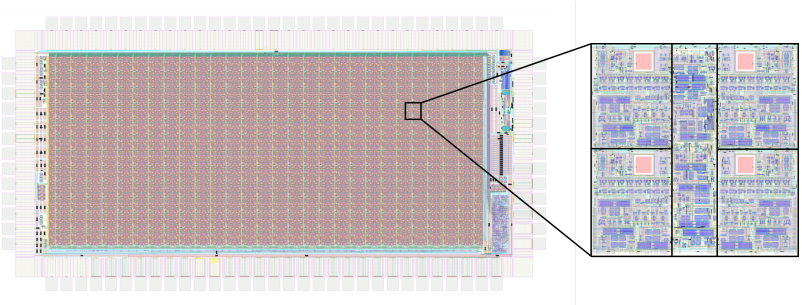


Figure 1: Chip layout.

Evaluation of Architectures for FPGA-Implementation of High-Resolution TDCs

M. Parsakordasiabi, I. Vornicu, R. Carmona-Galan, and A. Rodriguez-Vazquez
Instituto de Microelectronica de Sevilla (IMSE-CNM), CSIC-Univ. Sevilla, Spain

Abstract: Time-to-digital converters (TDCs) are a central component in systems based on time-delay assessment. The principal characteristics to be sought for in a TDC are high resolution, long time range, linearity and low power consumption. Besides, field-programmable gate arrays (FPGAs) represent an interesting option to explore fully-digital TDC architectures, because of their flexibility, shorter development time and lower implementation cost than ASICs. They are reconfigurable and usually built on the finest silicon technologies. The purpose of this work is to identify the different architectures that lead to high-resolution TDCs on FPGA, and to compare them in terms of the appropriate figures of merit. The most extended method to cover a long time interval while preserving a high time resolution is to combine a coarse counter with a fine time interpolator. Two techniques have been widely used to implement the interpolator, namely a tapped delay line (TDL) and a multiple-phase clock interpolator. Exploiting fast carry chains present in most modern FPGAs, sub-clock-period resolution have been achieved, down to tens of picoseconds. Other important aspects of the TDC design are the thermometer-to-binary encoder, the minimization of the clock skew, the analysis of the influence of voltage and temperature changes and bin-width calibration. Accordingly, we report an analysis of the different TDC architectures on FPGA based on their performance characteristics.

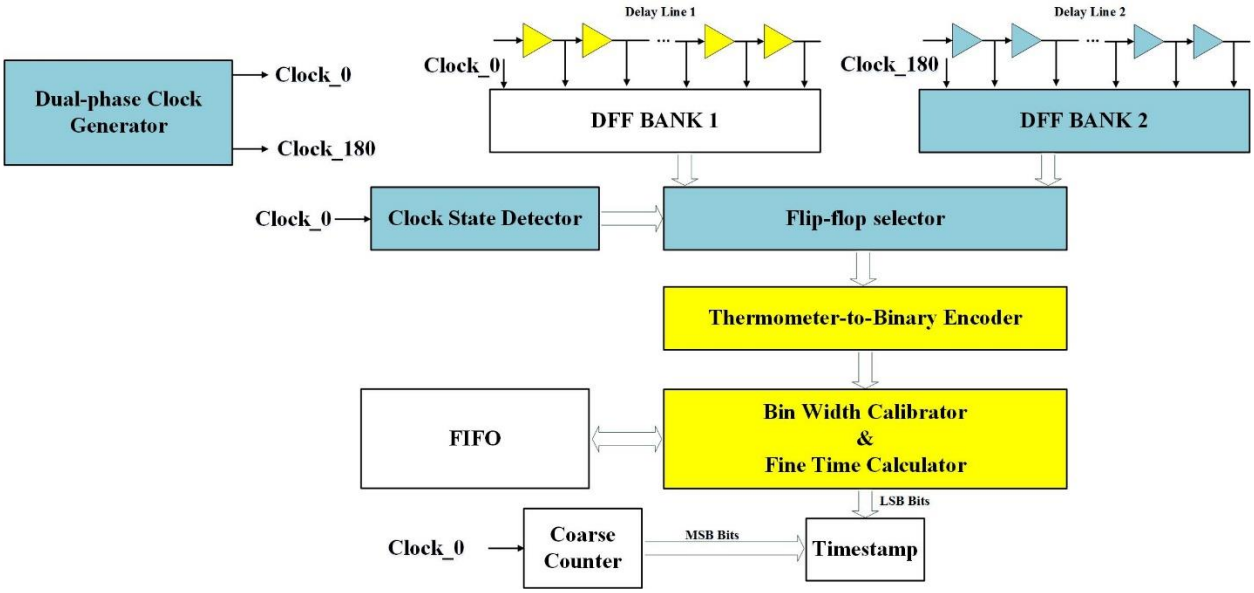


Fig. 1. Block diagram of a FPGA-based TDC architecture with TDLs and multi-phase clock

Acknowledgments: This work was supported by EU H2020 MSCA through Project ACHIEVE-ITN (Grant No 765866), by the Spanish MINECO and European Region Development Fund (ERDF/FEDER) through Project RTI2018-097088-B-C31 and by the US Office of Naval Research through Grant No. N00014-19-1-2156.

OpenDenoising : An Open Benchmark for Image Denoising Methods

Florian LEMARCHAND [1], Eduardo FERNANDES MONTESUMA[1],
Maxime PELCAT[1,2] and Erwan NOGUES[1,3]

[1] Univ. Rennes, INSA Rennes, CNRS, IETR - UMR 6164
[2] Institut Pascal, Clermont-Ferrand, UMR CNRS 6602
[3] DGA-MI

Abstract

The talk presents the on-going development of an open image denoising benchmark. This benchmark aims to provide an accessible interface to compare denoising methods and especially statistical ones. The architecture has been thought in a tunable manner to allow any user to use its algorithms, datasets, metrics. Several state-of-the-art methods are included in the benchmark to avoid re-training them. The benchmark currently supports native Python, Keras/Tensorflow and Matlab algorithm implementations. While originally built for methods addressing image denoising, the support of videos is in progress.

Image interpolation techniques

S.BOUKHTACHE, F.BERRY, M.GREDIAC, B.BLAYSAT

(Institut pascal, Clermont-Ferrand)

Abstract

Image interpolation is one of the fundamental operations used in digital image processing. It is widely used and required by many applications such as resolution enhancement, image resizing, geometric transformation, or in multiple domains such as displacement estimation with digital image correlation. Since the ideal interpolation kernel “sinc-function” is unlimited, many finite size interpolation kernels have been proposed and presented in the literature.

In this presentation, the most commonly used interpolation algorithms are surveyed. These algorithms are implemented on FPGA then compared in terms of interpolation quality (MSE & PSNR) and hardware resources consumptions (Adders, multipliers, registers, LUTs and ALMs).

Software architecture for cooperative sensing

Patrick Heyer Wollenberg, Chengjin Lyu, Ljiljana Platisa,
Bart Goossens, and Wilfried Philips

TELIN-IPI, Ghent University - imec, 9000 Gent, Belgium

Abstract: One of the main requirements for multi-sensor cooperative tracking, is the need to simultaneously capture and process information provided by a wide range of sensors, and the additional information obtained during the processing. This work presents a novel architecture that uses a shared memory scheme, that communicates a series of sensing and processing plugins being executed in parallel, each on its own independent thread. The system presented consists of four main components: *1)* The *shared memory* part, that uses a blackboard pattern that relies on a central index of pointers that contain the information to be shared, this information can be variables, objects, and even functions. 2*)* The *plugin manager*, using a singleton pattern to load, initialize, configure, and execute the plugins. *3)* The *shared logger* centralizes the system log by writing the debug and execution intimation from the system and the plugins to a central synchronized location. *4)* The *plugins* are the main part of the cooperative system, providing the necessary framework to capture, process, display, and record the information provided by sensors. The plugins of the system are designed to be easily implemented based on existing code by maintaining the external library dependencies of the architecture to a minimum, *(pthreads and c++x11) and platform independent (tested on Windows 7,8, vista, X, linux).* In this work we present the main design of the proposed architecture, the use cases it has been tested on, and the current and future work where it will be used.

Acknowledgments: This work was supported by EU H2020 MSCA through Project ACHIEVE-ITN (Grant No 765866).

Clip-level Feature Aggregation for Video-based Person Re-Identification

Chengjin Lyu, Patrick Heyer Wollenberg, Ljiljana Platisa, Bart Goossens, and Wilfried Philips
TELIN-IPI, Ghent University - imec, 9000 Gent, Belgium

Abstract: In the task of video-based person re-identification, sequence-level features of query and gallery persons are compared to search the best matching. Generally, due to the memory limitation of a single GPU, frame-level features are aggregated together using a temporal modeling method to generate clip-level features, instead of a sequence-level representation. Although these clip-level features have achieved impressive results, the importance of clip-level feature aggregation is still lack of study. In this paper, we investigate the aggregation of clip-level features in video-based person re-identification. Specifically, a simple but useful clip-level feature aggregation method is proposed, which consists of two parts, i.e., Average Aggregation Strategy (AAS) and Raw Feature Utilization (RFU). The experimental results demonstrate that this method can boost the performance of existing models. In particular, with the help of our clip-level feature aggregation method, we achieve 87.7% rank-1 and 82.3% mAP on MARS dataset without any post-processing procedure, which outperforms the existing state-of-the-art methods

Acknowledgments: This work was supported by EU H2020 MSCA through Project ACHIEVE-ITN (Grant No 765866).

Learning Domain and Pose Invariant Features between Pairs of Cameras for Person Re-Identification

Asad Munir, Christian Micheloni
University of Udine, Udine, Italy

Abstract: Person Re-identification (re-id) is a challenging task that aims to retrieve the given person's images from an image dataset. The key issues for learning robust person features are domain and pose variations in person images. Domain variations usually occur due to various environments inside the fields of view of different cameras (e.g. indoor and outdoor cameras), while pose variations are caused by the change in viewpoint of a person. Existing works deal this problem by considering either domain variations or pose variations in the dataset. In the proposed work, we address both these variations in a single model based on Generative Adversarial network (GAN) named as Domain and Pose Invariant GAN (DPI-GAN). This framework uses a CycleGAN and makes its generators conditioned on the given pose. The purpose of this network is to generate new images from one camera domain to another camera domain with new poses for every pair of cameras. These generated images along with original images are used to learn new deep re-id features free of domain variations and pose variations between cameras and person images respectively. This work is in progress and our approach is shown in Fig 1.

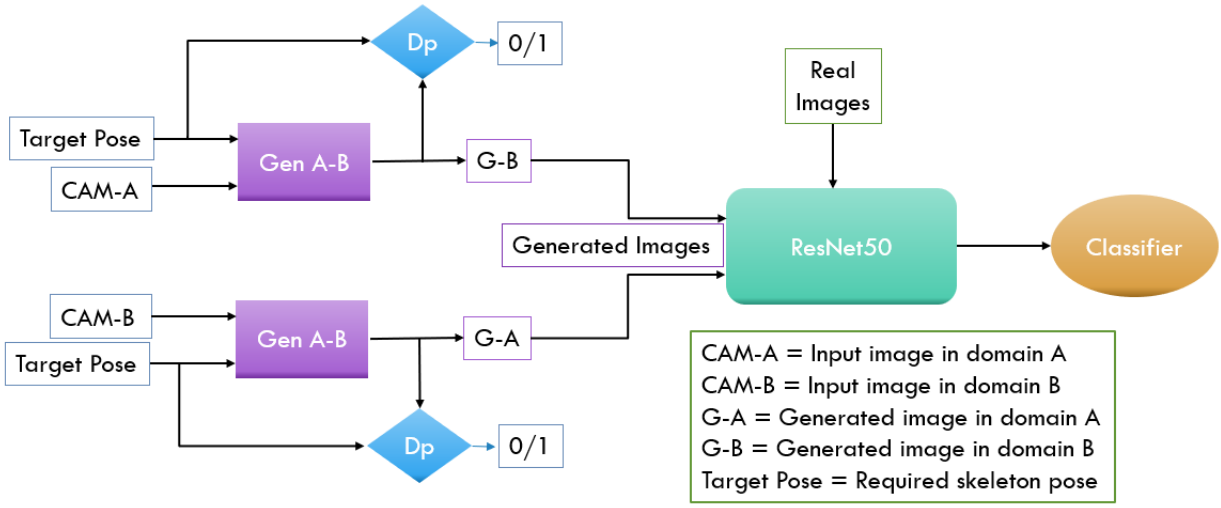


Fig. 1. Overview of our framework

Acknowledgments: This work was supported by EU H2020 MSCA through Project ACHIEVE-ITN (Grant No 765866)

Subject: Data and Task partition techniques and evaluation metrics for Deep Learning on FPGA-GPU Embedded Heterogeneous Architectures

As part of **Overview on ITN ACHIEVE PhD works on advanced hardware/software components for integrated/embedded vision systems.**

by Walther CARBALLO-HERNÁNDEZ

Dedicated and specialized hardware architectures covering Deep Learning inference deployment have been widely adopted in distributed data centers with promising results. However, most of these ad-hoc partitioning techniques do not take into consideration some limitations on specific device resources which are especially crucial on embedded systems. In this work we propose a data-oriented and task-oriented partitioning and evaluation methodology based on power consumption metrics and low bandwidth communication preserving latency and accuracy on specific models. As experimental results, we chose an Heterogenous Architecture for embedded vision applications with a Nvidia Jetson TX2 (SoM CPU+GPU) and an Intel Altera Cyclone10GX (FPGA), both based on a Development Kits. As part of the project, we propose an early stage embedded smart camera architecture design that includes both SoMs on a single platform and multiple power evaluation IC for research purposes and multiple setups. We also show a case study for layer task and data partitioning and their impact in feature transmission between both devices for first layers of YoloV2 object detection with 36.5% of data division on the FPGA side and 63.5% on GPU for deeper layers.

Hydra

A general-purpose multi view smart camera and its application to object recognition with an efficient convolutional neural network

Jonathan Bonnard[1], Kamel Abdelouahab[1], Maxime Pelcat[2], François Berry[1]

Multi-view imaging is the process of combining several views of the same scene.
In this way, it is possible to take advantage of the additional information to reconstruct 3D objects, generate depth maps or panoramas. However, these applications have a significant cost in terms of computations and represent a major bottleneck for most of the conventional electronic systems such as desktop processors. This downside is even more noticeable when targeting deep learning-based algorithms mainly involving convolutional neural networks (CNN).

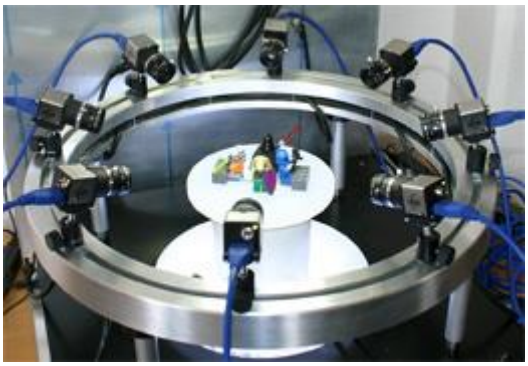
This thesis addresses the problem with an efficient multi-view smart camera able to concurrently process multiple views. Technically, the camera splits the acquisition into two distinct components:

- On one hand, each “camera head” embeds a FPGA to provide a pre-processing service at the nearest of the image sensors.
- On the other hand, a main “camera body” drives the communications, ensures the synchronization, supplies powers and gathers data for additional computations.

To highlight the effectiveness of this multi-view system, this work also provides an alternative solution to the usually complex task of object recognition with a CNN. While numerous contributions proposed to increase recognition rate with several state-of-the-art neural networks processed on each of the view perspectives, our system uses a severely downsized CNN deployed across the two parts of the system.
This partitioning scheme has two advantages:

- First, it takes advantage of the supplementary information acquired to increase the classification accuracy of CNNs, and even counteracts the accuracy loss due to pruning or layer removal.
- Second, the overall workload is split between “camera heads” and the central node

Preliminary results show a significant workload reduction to the AlexNet CNN and a similar classification accuracy in a four “camera heads” setup on the modelNet40 dataset.



Credits: https://en.ids-imaging.com/news-article/en_multicamera_system.html

[1] Institut Pascal, Université Clermont Auvergne, Clermont Auvergne

[2] Institut d’Electronique et de télécommunications de Rennes, INSA Rennes

3D sensing applications and techniques

A. Ferrario
IMASENIC S.L., Barcelona, Spain

Abstract: 3D imaging has been around for decades but an increased demand for complex imaging together with technological evolution have brought it to expand over a very broad spectrum of applications. One of the most important ones is 3D detection for medical diagnosis and surgery: the possibility to successfully accomplish very delicate and rare operations has been made feasible thanks to the ability to reproduce internal organs which can be used as trainers before performing real operations. To follow, depth sensing has gained importance in the automotive environment: car crash avoidance, driver's physical conditions monitoring and autonomous driving are now leading characters for customers which aim to reduce the death count due to vehicle accidents. Another important application is the one of security and data protection: face recognition together with the well-known fingerprint technique constitute nowadays a solid method to protect people from unwanted stealing of personal data. A considerable amount of different fields covering smaller spheres of interests also make extensive use of 3D sensing, videogames as an example. Several solutions for depth detection have been developed resulting in some techniques being more suitable than others depending on the application: parameters such as technologies, resolution, precision, robustness, range, computational and production cost will be discussed and compared for different techniques in an attempt of a state-of-the-art 3D detection techniques summary.