論　文　の　要　旨

題　目　　Study on Power-Efficient Acceleration Coprocessors in CMOS Technology for Real-time Embedded Machine Learning and Vision
（リアルタイム組込み機械学習及びビジョンのための CMOS 技術における電力効率の良い加速コプロセッサに関する研究 ）

氏　名　　張　湘煜

Humans can detect and identify a multitude of objects in a scene with little effort, despite the wide variability in appearance of outdoor scenes, such as complex backgrounds, different poses and illumination conditions. However, it is still challenging for a computer to recognize objects in an image or video sequence. Many efforts have been made to solve the task over multiple decades. Object recognition plays a key role in various fields, including smart vehicle technologies such as advanced driver assistance systems (ADAS), advanced human computer interfaces, robotics, surveillance, security, and intelligent transportation systems. For example, the ADAS uses the images acquired from a camera mounted on the vehicle, detecting the pedestrians and vehicles in images. The recognized results can be utilized to prevent accidents.

Pedestrian detection is a challenging work because of various clothes, changing lighting conditions, viewpoints, and a wide range of people's positions and sizes. Recent studies have shown that the sliding-window based methods can overcome these obstacles and obtain improved recognition performances. A feature descriptor extracts the representative data of an image, improving perception of the surrounding environment and transforming the sensed signals into a suitable data format required for the subsequent recognition processing. The histogram of orientation gradients (HOG) algorithm was proposed to distinguish pedestrians in images, obtaining good recognition results through extracting a robust feature set based on gradients. HOG features have been widely applied in image classification and scene understanding tasks.

However, the huge computational complexity of the HOG descriptor remains as a problem for the processing speed. An image has plenty of the sliding windows, and there are several thousand dimensional features in each sliding window. The enormous complexity of computation makes its application almost impossible for real-time processing. Studies have shown that the HOG descriptor can obtain good recognition performance and processing

speed in a PC environment. However, the vehicle and robot systems, which use an embedded processor platform, have no access to the same computing resources as PC environments. The feature space is too large for embedded systems and the processing power of an embedded platform is much lower than that of a PC platform, making it hard for systems based on embedded platforms to recognize objects in real-time.

In addition to recognition, learning internal representations of the perceived environment is essential because it bridges the gap between the representations of the object and the data needed by the computer to perform its task. Learning vector quantization (LVQ) neural networks have been successfully used for a broad range of technical applications, such as image compression and object recognition. In the literature, the LVQ was implemented in software off-line on computer systems or embedded processors. Unfortunately, the software-based approaches cannot deliver reasonable performance for online learning due to the high cost of computational requirements.

Hardware-based accelerators for computer vision can meet the low power and real-time processing requirements for mobile devices because the accelerator architecture can be tailored to specific applications and can be massively parallelized. A high energy efficiency and high throughput are the advantages over software-based solutions and general-purpose hardware.

In this thesis, I present two application specific integrated circuits (ASICs), aiming at low power, portable, and real-time applications and markets, such as Advanced Driver Assistance System (ADAS), robotics, drones, or mobile phones. One is an inference system designed to extract features from images using the HOG feature extractor and to reason based of the results of a nearest-neighbor-search (NNS) classifier. The other is an on-chip learning system aimed to learn from the images, featuring an LVQ neural network. The ability to reason and the ability to learn are the two major capabilities associated with these systems.

A) The inference system combines the feature extraction and dimension reduction in an intermediate step using partial-least-squares-regression in order to avoid the curse of high dimensionality. The design reduces the redundancy in original feature vectors, converting high-dimensional feature-vectors into low-dimensional feature-vectors. The following NNS works on feature vectors in a reduced-dimension space. The developed

hardware-oriented algorithm exploits the cell-based scan strategy which enables image-sensor synchronization and extraction-speed acceleration. Furthermore, buffers for image frames or integral images are avoided. The fabricated test chip in 180 nm CMOS technology achieves fast processing speed and large flexibility for different image resolutions with substantially reduced hardware cost and energy consumption. For the application example of XGA (1024 × 768) resolution videos, HOG-feature vectors can be extracted at 120 MHz operating frequency with a maximum frame rate of 122 fps. An improved version was fabricated in 65 nm CMOS technology which can process XGA (1024 × 768, 30 fps) video in real time, achieving 50 MHz feature extraction and 200 MHz classification, with energy consumption of 906 pJ/pixel. Detection accuracy can be improved using complementary features in addition to the HOG feature, at the cost of an extra 40% power consumption, 64% area requirement, and 53% memory size.

B) The on-chip learning system is based on a modular and reconfigurable pipeline architecture (MRPA) for LVQ. The MRPA consists of dynamically reconfigurable modules and realizes a run-time and on-chip configuration for recognition and learning. The developed architectures enable to speed up system development time and to provide better performance. The design effectively utilizes the available memory of the given hardware resources. Prototype fabrication in 65-nm CMOS technology verifies high integration density and memory-utilization efficiency, good performance, and considerable flexibility in vector dimensionality, number of weight-vectors, and adaption strategies. Compared with embedded microprocessors, which rely on single-instruction-multiple-data (SIMD) processing, the developed MRPA-prototype increases the performance of both recognition and learning operations. The achieved improvements amount to approximately factors 40 and 101 on the well-established performance metrics of million connections per second (MCPS) for recognition and of million connection updates per second (MCUPS) for learning, respectively.

The prototype ASIC consumes 21.5 mW working at 150 MHz and 1 V voltage, with 2.14 mm^2 area overhead in 65nm CMOS technology. A small accuracy loss mainly comes from the truncation operation of the fixed-point operation, resulting in a quite small peak signal-to-noise ratio (PSNR) loss of 0.128 dB. The applied pipeline reconfiguration leads to a reduction in computation time and high efficiency for integration density. The applied modularity contributes to easy scalability in a both upward- and downward-compatible fashion. Additionally, the introduced shared memory-pool increases the flexibility for both

the dimensionality and the number of weight vectors. Further, an implemented parameterization for system configuration adds flexibility to the choice of adaption strategies in different applications. Robustness Analysis of Vibration Features Against Appearance Changes in High-Frame-Rate Videos

I hope after several decades, the generation takes the computer vision for granted. With the futuristic mobile device, if they see some bug or a little crab on the beach, they will just assume that it has always been the case that they can just snap a photo and that system will tell them what it is and everything they could want to know about it. Two feasible directions to further improve the system performance are listed below.

A) Enhancing Pedestrian Detection Accuracy

   Our group has tried to include complementary feature descriptor (Haar-like) for HOG to improve the accuracy. One future direction would be integrate more sliding-window based feature descriptors in one chip and vote on the basis of the detection results from all the descriptors. Besides, the current version has fixed primitive-size (cell, window, sliding step). So another direction is parametrize the primitive-size and the image size. In addition, enabling the learning function into the detection circuits can also improve the accuracy.

B) Embedding with Lane Detection

   Lane detection has a great contribution on traffic-safety, as it is the major contributor to a lane departure warning (LDW) system, which is a basic and necessary part for an Advanced Driver Assistant System (ADAS). LDW is a system that uses the information from lane detection to warn the driver of lane departure. Then, the driver can correct the route to avoid that potential accidents happen. Due to its significance, many researchers pay attention and work on the study of LDW system for the advances in self-driving technologies. Accordingly, a real-time, robust and accurate lane-detection method is necessary for vehicle navigation. Combing the pedestrian detection and lane detection will result in a comprehensive result.