A 69mW 140-meter/60fps and 60-meter/300fps Intelligent Vision SoC for Versatile Automotive Applications

Yi-Min Tsai, Tien-Ju Yang, Chih-Chung Tsai, Keng-Yen Huang, and Liang-Gee Chen
DSP/IC Design Lab, Graduate Institute of Electronics Engineering, National Taiwan University, Taipei, Taiwan
Email: ymtsai@video.ee.ntu.edu.tw

Abstract
A machine-learning based intelligent vision SoC implemented on a 9.3 mm² die in a 40nm CMOS process is presented. The architecture realizes 140 meters active distance at 60fps and 60 meters at 300fps under Quad-VGA (1280×960) resolution while maintaining above 90% detection rate for versatile automotive applications. The system supports 64 object tracking and prediction. It raises 1.62 times improvement on power efficiency and at least 1.79× increase on frame rate with the proposed knowledge-based tracking processor. The chip achieves 354.2fps/W and 3.01TOPS/W power efficiency with 69mW average power consumption.

Introduction
For versatile automotive applications, several issues are considered while developing a reliable vision system, including active distance, driving speed. Active distance is defined as the maximum detecting distance to objects. According to the driving safety rules, it requires approximate 50 meters safety distance while driving at 100 Km/hr, so the corresponding active distance should be larger than 50 meters. Besides, the average response time of human is about 0.1s, so the time step between successive frames should be smaller than 0.1s, implying at least 10fps frame rate is required. The criteria suggest necessity of long active distance and high processing capability for high speed driving conditions. Another requirement is to actively locate and monitor surrounding objects, such as overtaking vehicles, preceding vehicles, or moving pedestrians, for further behavior analysis. For most applications, the angular distance to objects. According to the driving safety rules, it requires driving speed. Active distance is defined as the maximum detecting distance to objects. According to the driving safety rules, it requires approximate 50 meters safety distance while driving at 100 Km/hr, so the corresponding active distance should be larger than 50 meters. Besides, the average response time of human is about 0.1s, so the time step between successive frames should be smaller than 0.1s, implying at least 10fps frame rate is required. The criteria suggest necessity of long active distance and high processing capability for high speed driving conditions. Another requirement is to actively locate and monitor surrounding objects, such as overtaking vehicles, preceding vehicles, or moving pedestrians, for further behavior analysis. For most applications, the angular distance to objects. According to the driving safety rules, it requires driving speed. Active distance is defined as the maximum detecting distance to objects. According to the driving safety rules, it requires approximate 50 meters safety distance while driving at 100 Km/hr, so the corresponding active distance should be larger than 50 meters. Besides, the average response time of human is about 0.1s, so the time step between successive frames should be smaller than 0.1s, implying at least 10fps frame rate is required. The criteria suggest necessity of long active distance and high processing capability for high speed driving conditions. Another requirement is to actively locate and monitor surrounding objects, such as overtaking vehicles, preceding vehicles, or moving pedestrians, for further behavior analysis. For most applications, the angular distance to objects.

Intelligent Vision Automotive System
We develop a cooperative detect-and-track vision system to recognize objects and analyze dynamics of them (Fig. 1). We adopt the machine-learning based framework, AdaBoost with Haar-like features [6], for on-road object detection in multi-scale resolution for various active distance supporting. The active distance is related to the corresponding resolution (layer) in the image pyramid, given that the visual recognition system guarantees an accurate recognition result under that resolution. A large amount of Haar-like features are required to train strong classifiers to provide sufficient recognition quality and speed. The detecting window scans horizontal ROI specified dynamically based on the feed-backed object locations to reduce computation. In contrast with SIFT-based systems [2, 3, 5], Haar-like based systems have preferable discriminative properties for long-range object processing. We also bring up a knowledge-based Kalman object tracking mechanism considering spatial shape cues, especially for symmetrical appearances, to reveal dominant physical characteristics (i.e., object width, relative distance and speed), which provides further adaptive control of the system.

Proposed SoC Architecture
The proposed architecture is shown in Fig. 2. A Learning-based Detection Processor (LDP) and a Knowledge-based Tracking Processor (KTP) are controlled by a 32-bit reconfigurable RISC, Decision Processor (DP). 64-bit data bus is applied to access gray-scale image pyramid and color-scale (a*b*) data from an off-chip DRAM for LDP and for KTP respectively. 54-bit Wishbone bus is employed for DP. Two 16-bit data/instructions are individually concatenated into 32 bits in successive two cycles to reduce DP bus bandwidth. DP executes low-complexity programs, including resolution selection, ROI adjustment, and relative distance/speed estimation. It also copes with result transmission between LDP and KTP. The architecture supports detecting window width (w_s) of 20 and 32 pixels with 4 diverse aspect ratios, and maximum ROI height (h_ROI) of 120 pixels under 1280x960 maximum resolution which gives a specification of 140-meter active distance and 90-degree field-of-view.

Fig. 1 Proposed intelligent vision automotive system.

A. Learning-based Detection Processor
Fig. 3 shows Recognition Engine (RE) and Mesh Unit (MU) of LDP. For object detection, the windows (candidates) are verified by trained classifiers composed of cascaded stages. In each stage, Haar-like feature values (HFV) are calculated to compare to weak classifier thresholds (C_t, c) to accumulate weighting values for comparing with a stage threshold (S_h, c). Since classifiers are accessed frequently, to reduce bandwidth and data accesses, features and thresholds are stored in a 4-bank on-chip classifier memory with capacity of 512 features constructing 20 stages. 4 classifier memories are designed to allow various window aspect ratios or object types. Noting that a Haar-like feature is structured as two or three adjacent rectangles that can be represented by unit rectangles, LDP stores compressed features and decode them on-the-fly. This achieves 43% classifier memory reduction (from 21KB to 12KB). RE decodes 4 features into 48 positions and computes 4 HFVs in parallel. It requires 2 to 16 processing cycles for each window depending on the number of features in a stage. The former stage contains fewer features and approximate 70% windows are early rejected by the first and second stages, leading to 73% reduction in feature calculation time. Through RE, LDP reaches 0.28G feature's processing capability. The feature calculation involves integral pixel (ip) and square integral pixel (sip) techniques [6]. 16-bank integral pixel memory and 8-bank square integral pixel memory are allocated for ip and sip storing respectively. It holds 87.5% data reuse while shifting the window. Integral Pixel Calculator (IPC) computes at most 16 new ips or updates 16 existing ips of integral pixel memory, which is able to reduce 91% input bandwidth (Fig. 4). Square Integral Pixel Calculator (SIPC) is similar to IPC except it recovers 16 pixel values from 32 ips as input pixels for calculating 16 sips. MU comprises a 2D shift-register array with row-column multiplexers for irregular random ip access and three 1D shift-register queues for sip access. It receives the decoded 48 positions each cycle and achieves maximum 9.6G ips/s data throughput. The partial ips for the next window is loaded to MU when RE processes the current window, which increases LDP utilization to accelerate the work scheduling. Lastly, intra-clustering and inter-clustering are performed to group results belonging to the same object and produce a confidence value (cv).

B. Knowledge-based Tracking Processor
Fig. 5 illustrates KTP. The tracking routine consists of shape extraction, symmetry calculation, and state estimation. Shape Engine (SHE) transforms pixel data into run-length representations for extracting intact object boundaries. It encodes the binarized data generated via pixel-wise comparison into length data and performs morphological operations to unite scattered segments (lengths) with previous encoded segments. 8 length values are sufficient to describe a row in the bounding box, implying an 8-bit KTP memory for efficient length storing. Segments of all rows are rotated in a circular shift-register and the one at the head position is encoded. This reduces 90% data access of KTP memory and

978-1-4673-0849-6/12/$31.00 ©2012 IEEE 2012 Symposium on VLSI Circuits Digest of Technical Papers 152
shortens encoding latency. Symmetry Engine (SYE) performs object refinement and calculates symmetry value (sv), defined as the sum of subtractions of length values of rows in the bonding box, as the matching basis for the following state estimation. A range of shift offsets (so) are applied to determine the most symmetrical axis with the smallest sv. SYE contains 32 PE s for computing multiple rows in parallel and requires at most 8 cycles/so for an object. State Engine (STE) performs Kalman prediction describing object dynamics. Maximum 64 predictors are supported. Active distance (m) (2.6×2.6 mm²) and 1 bit/2.5V voltage level. It dissipates 69mW average power at 220MHz with 0.9V/2.5V supply. Moreover, system latency less than 30ns is promised. The architecture realizes 140 meters active distance at 60fps and 60 meters at 300fps with above 90% detection rate accompanying less than 4% false alarm rate for both cases.

Chip Implementation Results

The chip is implemented in TSMC 40nm 1P10M CMOS process. Fig. 6 summarizes chip performance. Clock gating and multi-threshold voltage for logic cells are employed to reduce total power consumption. The die size is 3.0×3.1mm² (2.6×2.6 mm² for core) including IO and bonding pads. It dissipates 69mW average power at 220MHz with 0.9/2.5V core/IO voltage. Fig. 7 shows the comparison between several recognition works and the proposed architecture. 3.01TOPS/W power efficiency and 55.6GOPS/mm² area efficiency are achieved. For Haar-like object detection, the processing efficiency is 0.327fps/MHz under VGA resolution with 3.6x to 8.8x outperformance. Moreover, system latency less than 30ns is promised. The architecture realizes 140 meters active distance at 60fps and 60 meters at 300fps with above 90% detection rate accompanying less than 4% false alarm rate for both cases.

References