A 52mW Full HD 160-Degree Object Viewpoint Recognition SoC with Visual Vocabulary Processor for Wearable Vision Applications

Yu-Chi Su, Keng-Yen Huang, Tse-Wei Chen, Yi-Min Tsai, Shao-Yi Chien, and Liang-Gee Chen
Graduate Institute of Electronics Engineering and Department of Electrical Engineering
National Taiwan University, Taipei, Taiwan
E-mal: steffi@video.ee.ntu.edu.tw

Abstract

A wearable 1920×1080 160-degree object viewpoint recognition SoC is realized on a 6.38mm^2 die with 65nm CMOS technology. This system focuses on enhancing the capability for wide viewpoint and long-distance recognition while reducing the computation of feature matching process. The recognition accuracy is improved from 29% to 94% under full HD resolution for a 50m-far traffic light compared with the performance under VGA (640×480). Object viewpoint prediction (OVP) supports 160-degree object viewpoint differences. 85% of power consumption and 75% of memory bandwidth are reduced via proposed visual vocabulary processor (VVP). 52mW power consumption with 25.9GOPS/mm^2 area efficiency is achieved.

Introduction

Recently, mobile vision technologies, such as augmented reality, robot vision, and visually-impaired electronic aids, have been developed to assist people and make our lives more convenient. However, in many circumstances especially when wearing these devices in outdoors, state-of-the-art techniques show limited performance. We attribute this phenomenon into three major causes: (1) difficulty in detecting long-distance or small-sized objects, (2) poor recognition accuracy under large object viewpoint variation and dramatic camera ego-motions and (3) high power consumption due to the complex computation and frequent memory access.

In this paper, we propose a full HD 160-degree (80° for one side) object viewpoint recognition SoC as shown in Fig. 1. For overcoming the above shortages, three prominent characteristics are introduced in our system. Firstly, to recognize objects at far distance or with small size, the proposed vision recognition system is designed for full HD resolution with 30fps. Higher resolution leads to better performance in recognizing an object occupying a small portion of an image as illustrated in Fig. 1. Secondly, OVP is proposed to allow 160-degree variation of object appearance. Through synthesizing predicted pose candidates of an object, the capability of viewpoint variation tolerance is significantly enhanced without feeding extra images into the database. Lastly, VVP is designed to simplify the complicated computation. Existing object recognition systems [1-3] operate object recognition in feature matching stage and require frequent memory accesses. More memory access leads to higher power consumption that is critical in wearable applications. In this work, we advance the matching process from feature level to object level via VVP. It utilizes the conceptions of Bag-of-Words (BoW) object representation and the vocabulary tree [4] to characterize an object as a histogram vector. Instead of matching features that results in thousands of memory fetching, VVP only compares the histogram vector with memory access once to recognize an object. Combined with the above three distinguished characteristics, the proposed recognition SoC achieves both high accuracy and power efficiency for wearable vision applications.

Wearable Recognition SoC Architecture

Fig. 2 shows the block diagram of the proposed visual recognition system. The whole system can be roughly divided into two levels, feature-level and object-level. For feature-level operation, the human-centered design (HCD) is the preprocessing stage for our system. It is composed of camera motion stabilization (CMS), OVP, and attention tracking (AT) engine. CMS computes camera motion by tracking static objects in video sequence and further stabilizes video by compensating severe translational camera ego-motion in each frame. In our system, OVP provides maximal 160-degree viewpoint of object appearance through predicting the possible poses from CMS. AT defines the ROI on the current frame from previous spatial information of the detected object. Feature detection and description modules, which implement SIFT algorithm [5], extract features from attention regions of the image. Feature matching processor performs conventional all-feature matching every thirty frames. For the rest of frames, the operating procedure is raised into the object-level processing. VVP handles feature voting within ROI to gather a cluster of features to represent an object as a histogram vector. The object histogram comparator in VVP compares the histogram vectors with referenced objects in the database to classify the categories of the detected target.

Detailed Circuits of VVP and HCD

To accelerate the speed of object matching, a massively-parallel architecture VVP is employed to achieve the real-time processing capability and the low-power-consumption requirement. In each stage of VVP, there are two distance processors, each of which includes 16 parallel processing elements (PE) and a tree-like adder for the calculations of the Euclidean distance as shown in Fig. 3. There are six stages in VVP architecture which is able to compute the 64-word representation of an input vector with a 16 dimensions/cycle throughput. The hierarchical memory, which contains 126 words in total, has 6 banks of memory to offer the data to the distance processors for the nearest-neighbor computations in each stage. The distance processors are connected to their corresponding banks of the hierarchical memory, and the max bandwidth of 343GB/s can be achieved when operating under the frequency 200MHz. It requires only 8 cycles to process a 128-dimensional vector, and more than 5 times of the bandwidth is saved compared to the non-binary-tree-based architecture.

Fig. 4 depicts the detail architecture of HCD. HCD consists of CMS, AT, and OVP modules. CMS contains a 128 parallel vector processing elements (VPE) cluster to compute confidential weighting for each feature as time elapses. A tree-based accumulator processes each feature within 14 cycles in pipeline manner and analyzes the global camera motion. The AT implements Generalized Hough Transform algorithm [6] with 8 parallel processors and 4 Hough-table voting (HV) mechanisms to group features of each object in the first frame and generates attention windows for the next fame. The subsequent 29 frames receive the input camera motion from CMS to predict attention windows for target objects. OVP is adopted to synthesize ROIs with multiple viewpoints of object poses according to the estimated viewpoint parameters from CMS. The SIFT processing module is functioned on the synthesized ROIs to extract features with variant viewpoint tolerance.

Implementation Results

Fig. 5 reveals the chip specification of the proposed system. The wearable recognition SoC is implemented in 65nm CMOS process with 6.38mm^2 including IO and bonding pad. For 1920×1080 with 30fps video sequences, only 52mW is consumed under 200MHz. Fig. 6(a) shows that the bandwidth of matching process is reduced by 97% through VVP processing. Above 94% recognition accuracy is accomplished within 80° of object viewpoint in one side (160° in total) as illustrated in Fig. 6(b). The peak performance is 165GOPS while the average power efficiency is 1.18TOPS/W. The area efficiency of 25.9GOPS/mm^2 is 2× better than the previous works [2][3]. Overall system comparison is listed in Fig. 7. Our proposed work exhibits remarkable performance in functionalities and efficiencies.

References

[1] Jinwook Oh et al., "A 1.2 mW On-Line Learning Mixed Mode Intelligent Inference Engine for Robust Object Recognition," in *IEEE Symposium on VLSI*, pp.17-18, 2010.

[2] J.-Y. Kim et al., "A 201.4GOPS 496mW Real-Time Multi-Object Recognition Processor with Bio-Inspired Neural Perception Engine," in *IEEE ISSCC*, pp.150-151, 2009.

[3] S. Lee et al., "A 345mW Heterogeneous Many-Core Processor with an Intelligent Inference Engine for Robust Object Recognition," in *IEEE ISSCC*, pp.332-333, 2010.

[4] D. Nister and H. Stewenius, "Scalable Recognition with a Vocabulary Tree," in *IEEE CVPR*, pp.2161-2168, 2006.

[5] D.G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," International Journal of Computer Vision, vol.60, no 20, pp.91-110, 2004.

[6] O. Barinova et al., "On Detection of Multiple Object Instances Using Hough Transforms," in *IEEE CVPR*, pp.2233-2240, 2010.

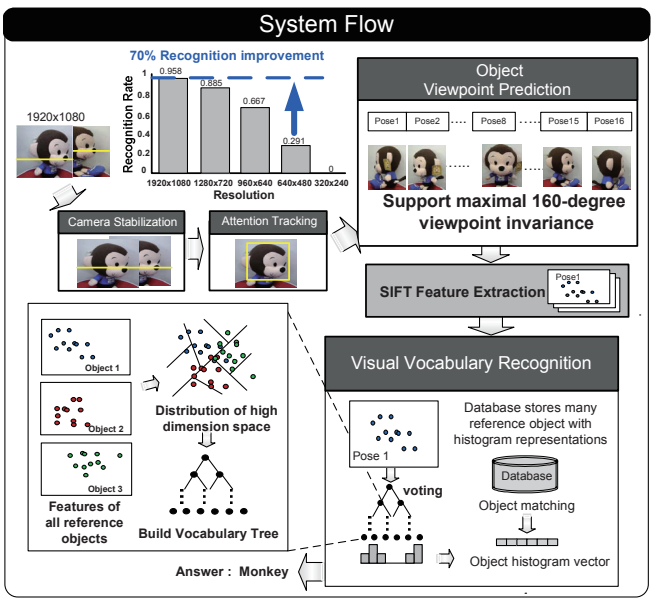

Fig. 1 System flow of the proposed wearable recognition SoC

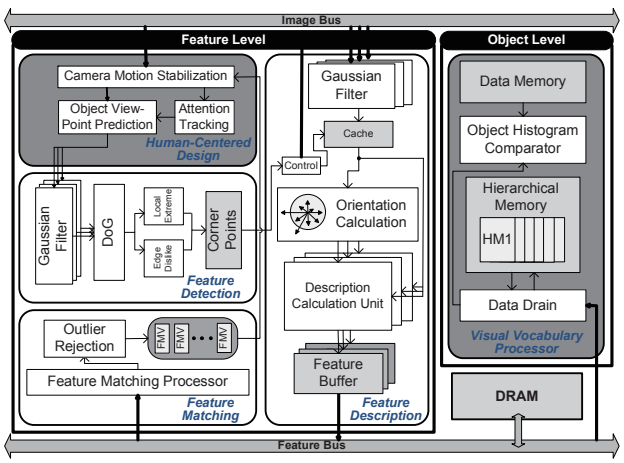

Fig. 2 System block diagram

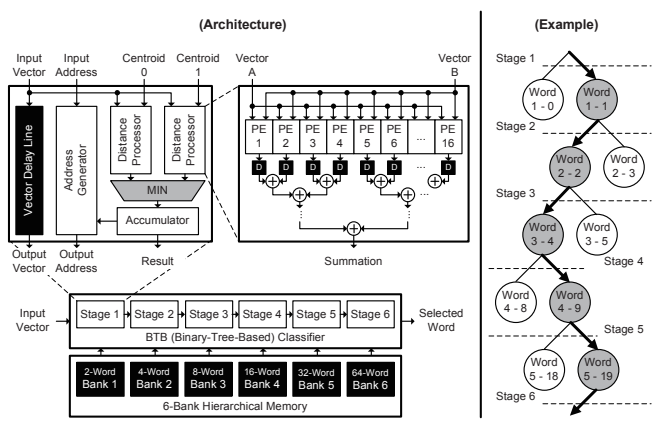

Fig. 3 Visual vocabulary processor architecture

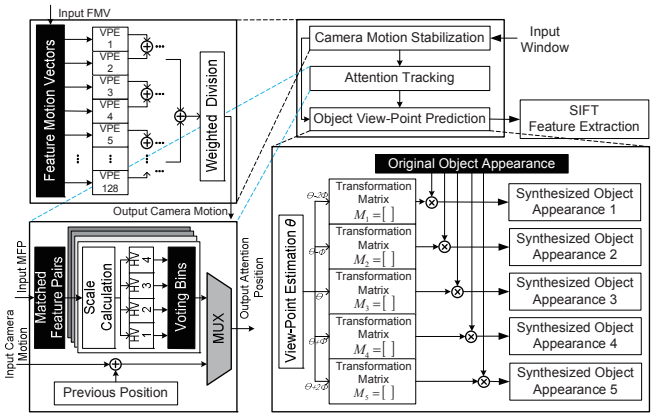

Fig. 4 Architecture of human-centered design

Item	Specification
Technology	TSMC 65nm 1P9M CMOS
Die Size	2.5mm x 2.6mm
Gates / SRAM	907.39K Gates / 40KB
Operating Freq.	200MHz
Power Supply	Core Power 1.0V
Power Consumption	Average 52.5mW
Peak Performance	164.95 GOPS
Power Efficiency	1.18 TOPS/W
Input Image	Full HD(1920x1080 30fps) Video Image

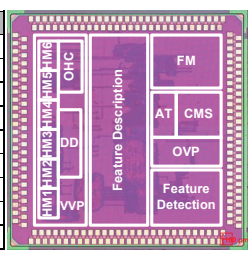

Fig. 5 Chip specification and photograph

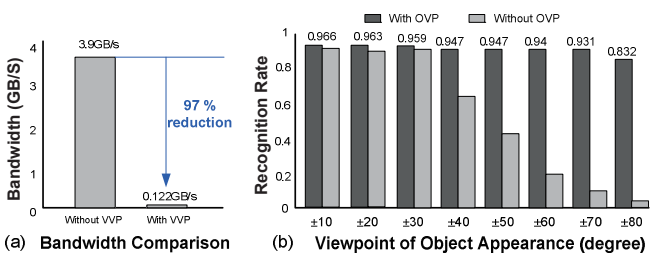

Fig. 6 Implementation results

	ISSCC 2009 [2]	ISSCC 2010 [3]	This work
Object Viewpoint Funcionality	Not Supported	100-degree	160-degree
Resolution	VGA(640x480)	VGA(640x480)	Full HD(1920x1080)
Average Power Consumption	496mW	345mW	52mW
Peak Power Consumption	695mW	704mW	198.4mW
Technology	0.13um	0.13um	65nm
Logic Gate Count	3.73M	2.92M	0.91M
On-Chip SRAM	396KB	612KB	40KB
Die Size	7.0mmx7.0mm	10.0mmx5.0mm	2.5mmx2.6mm

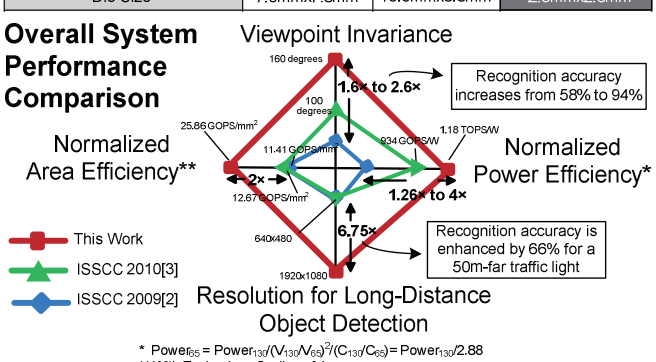

Fig. 7 Comparison with the previous works