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

Yu-Chi Su, *Student Member, IEEE*, Keng-Yen Huang, *Student Member, IEEE*, Tse-Wei Chen, *Member, IEEE*, Yi-Min Tsai, *Student Member, IEEE*, Shao-Yi Chien, *Member, IEEE*, and Liang-Gee Chen, *Fellow, IEEE*

***Abstract*—A 1920 × 1080 160° object viewpoint recognition system-on-chip (SoC) is presented in this paper. The SoC design is dedicated to wearable vision applications, and we address several crucial issues including the low recognition accuracy due to the use of low resolution images and dramatic changes in object viewpoints, and the high power consumption caused by the complex computations in existing computer vision object recognition systems. The human-centered design (HCD) mechanism is proposed in order to maintain a high recognition rate in difficult situations. To overcome the degradation of accuracy when dramatic changes to the object viewpoint occur, the object viewpoint prediction (OVP) engine in the HCD provides 160° object viewpoint invariance by synthesizing various object poses from predicted object viewpoints. To achieve low power consumption, the visual vocabulary processor (VVP), which is based on bag-of-words (BoW) matching algorithm, is used to advance the matching stage from the feature-level to the object-level and results in a 97% reduction in the required memory bandwidth compared to previous recognition systems. Moreover, the matching efficiency of the VVP enables the system to support real-time full HD (1920 × 1080) processing, thereby improving the recognition rate for detecting a traffic light at a distance of 50 m to 95% compared to the 29% recognition rate for VGA (640 × 480) processing. The real-time 1920 × 1080 visual recognition chip is realized on a 6.38 mm² die with 65 nm CMOS technology. It achieves an average recognition rate of 94%, a power efficiency of 1.18 TOPS/W, and an area efficiency of 25.9 GOPS/mm² while only dissipating 52 mW at 1.0 V.**

***Index Terms*—Digital circuit, hardware architecture, multimedia processing, object recognition, real-time processing, system-on-chip (SoC), wearable applications.**

I. Introduction

ADVANCES in computer vision technologies have enabled the development of many innovative applications that were previously unfeasible, such as intelligent vehicles, robotic vision, interactive gaming, and wearable vision. Wearable vision, which can be used in augmented reality applications and electronic aids for the visually impaired, is one of the applications of computer vision technologies with great potential for assisting people in the future. In these applications, object recognition is a fundamental technology that is necessary to perceive objects and to establish the spatial relationship among detected objects. In general, an object recognition system contains two basic stages: feature extraction and feature matching. For feature extraction, the Scale Invariant Feature Transform (SIFT) [1] is widely used due to its scale, rotation and illumination invariance. However, extracting SIFT features is very time-consuming even for a VGA image [2]. During the feature matching stage, each extracted feature in the video frame requires a search of its nearest neighbors among a great number of object features stored in a database. Such frequent memory access results in degradation of the performance and high power consumption for the entire system. Because both stages of object recognition require considerable computational resources, it is difficult to realize a real-time object recognition system using modern PCs.

Many studies have been conducted with the aim of reducing the processing time of object recognition. In [3], SIFT feature extraction for a 320 × 240 video was accelerated by operating eight processors in parallel. Massively parallel single instruction multiple data (SIMD) processors were utilized to generate feature descriptors and to match these descriptors with those in a database [4], [5]. Through a tile-based approach, SIMD processors were further doubled to 16 processors to achieve a processing rate of 30 frame per second (fps) using VGA images [6]. In [7]–[9], FPGA implementations for object recognition were proposed. The systems proposed in [7] and [9] can achieve real-time object recognition performance with VGA video input. Recently, human-like top-down feedback attention was adopted in [10] to reduce the computational complexity of the object recognition process.

However, the crucial issues of reliability and usability have prevented the existing recognition systems from being widely adopted in wearable or other portable vision applications. The main limitations can be summarized as follows. First, there is poor recognition accuracy under large changes in the object viewpoint. In [11], local feature descriptors utilized in previous studies [2]–[10] were found to have a tolerance of 50° for changes in the object viewpoint, which was the highest of all tested descriptors. Because of this limitation, the results of ob-

Manuscript received September 04, 2011; revised November 23, 2011; accepted December 20, 2011. Date of publication February 23, 2012; date of current version March 28, 2012. This paper was approved by Guest Editor Makoto Nagata. This work was supported by TSMC and the National Science Council under Grant NSC 100-2221-E-002-248-MY3.

Y.-C. Su, Y.-M. Tsai, S.-Y. Chien, and L.-G. Chen are with National Taiwan University, Taipei City 100, Taiwan (e-mail: steffi@video.ee.ntu.edu.tw).

K.-Y. Huang is with MediaTek, Hsinchu 30078, Taiwan.

T.-W. Chen is with the VLSI Design and Education Center, University of Tokyo, Tokyo 113-0032, Japan.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSSC.2012.2185349

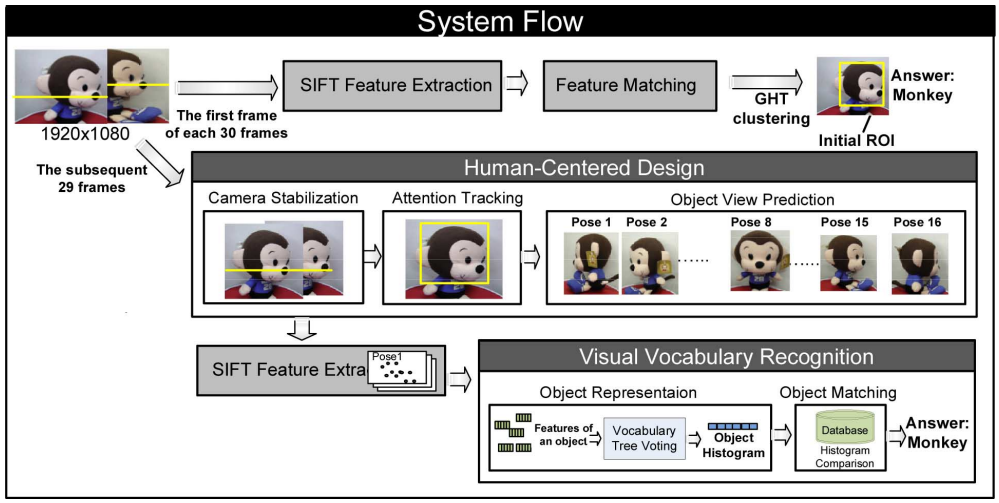

Fig. 1. System flow.

ject recognition can become unreliable. Second, low resolution inputs can cause difficulties in detecting small objects or objects at a distance. Due to the high computational cost for object recognition, previous studies [2]–[10] have aimed to recognize objects using low resolution video inputs, for example QVGA or VGA. Unfortunately, low resolution images are unable to provide enough detail to detect distant or small-sized objects. Third, there is typically a high power consumption due to the complex computational requirements of object recognition, especially the frequent memory accessing in the feature matching stage. The feature matching stage that performs highly frequent memory accessing is not only the performance bottleneck of the whole system but also consumes such a large amount of power. This situation becomes more serious when matching input features using a huge database containing millions of features. Despite focusing on the processing of QVGA or VGA resolution video inputs to avoid the high computation cost and power consumption, previous systems [2]–[9] still work for no longer than 8.8 h when supplied by a lithium battery.

To address the fundamental problems involved in object recognition, as mentioned above, we propose a full HD 160° (80° for one side) object viewpoint recognition SoC for wearable vision applications as shown in Fig. 1 [12]. To improve the recognition accuracy in challenging environments and to achieve low power consumption for practical use, three novel characteristics are introduced in our system. First, the human-centered design (HCD) engine is proposed, which allows a 160° variation in the viewpoint of objects even with severe camera shake. The HCD engine contains a camera motion stabilization (CMS) module and an object view point prediction (OVP) module. Through stabilizing the input video and synthesizing predicted pose candidates of an object, the viewpoint variation tolerance is significantly enhanced to 160° (80° for each side) without requiring extra images to be fed into the database. Second, to recognize distant or small-sized objects, a visual recognition system is designed for processing full HD videos at 30 fps. As shown in Fig. 2(a), the traffic light, which is about 50 m from the camera, occupies a small area in the VGA resolution image. As the result in Fig. 2(b) illustrates, a higher resolution leads to better performance in recognizing an object that is 50 m away. In this figure, the recognition rate is defined as the number of correctly detected target objects with respect to the total number of target objects that are present in the video. Nevertheless, processing video input with larger resolution leads to a larger number of extracted features. This result causes higher power consumption when matching numerous features and requires frequent memory access. To solve this problem, a visual vocabulary processor (VVP) has been designed to reduce the number of times memory must be accessed during feature matching. We have advanced the matching process from feature-level to object-level via the VVP. The VVP utilizes a bag-of-words (BoW) [13], [14] representation and a vocabulary tree [15] to characterize an object as an object histogram. Instead of applying feature matching, which results in millions of memory fetching operations, VVP compares the histogram only once to recognize an object. As a result, the power consumption is greatly reduced. The proposed visual recognition SoC, which incorporates the above three distinguishing characteristics, is highly accurate, efficient, and has a low power consumption that makes it suitable for wearable vision applications.

This paper is organized as follows. Section II introduces the algorithm of the visual recognition system. Section III illustrates the system architecture and detailed designs of each proposed module. The VLSI implementation of the proposed work and validation results are presented in Section IV. Finally, Section V concludes the work.

II. Algorithm

A. System Flow

Fig. 1 shows the system flow of the proposed visual recognition SoC. Once every 30 frames, the system runs full-frame feature-level matching based on traditional object recognition approaches. This is called the fine-grained object recognition stage in this paper. In the subsequent 29 frames, object-level matching is performed to accelerate the matching stage, and this is referred to as the coarse-grained object tracking stage.

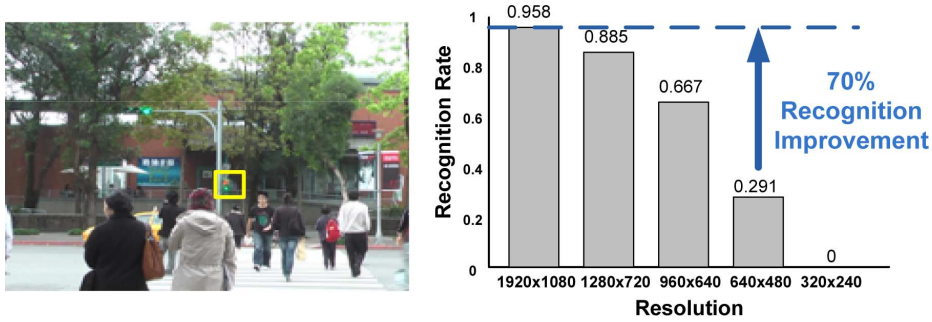

Fig. 2. (a) Distant object like the traffic light that occupies a small region in the image is difficult to be recognized. (b) The recognition rate versus the processing resolution.

During the fine-grained object recognition stage traditional full-frame recognition and matching is used to process the first frame of each 30 frames. First, the input frame is processed using a feature extraction method to generate features. In this work, the SIFT descriptor is employed to extract the local features of objects due to its invariance with respect to changes in scale, rotation, and illumination. The feature extraction step is divided into two phases: key-point detection and feature description. Key-point detection involves the scale-space extraction of extrema and is followed by the feature description step. The feature description step produces a descriptor of the detected key-points as 128-D feature vectors. After that, the generated features are processed by the feature matching stage that identifies matching pairs of features. Next, the generalize Hough transform (GHT) [16] is adopted to generate an initial region of interest (ROI) for each detected object. The GHT algorithm is a classic feature-level object recognition method that groups features corresponding to the same poses of an object. Each object in the database has a hash table that quantizes the object pose space by the values of four attributes, which are the x- and y-coordinates, the scale, and the rotation. The pose space is utilized to approximate a six degree-of-freedom (DoF) pose space for a 3-D object in the real world. After this step, initial ROIs representing the detected target objects are generated.

As the subsequent 29 frames are processed in the coarse-grained object tracking stage, the generated ROIs of the target objects are tracked from image to image. Two novel and innovative techniques are proposed to improve both the accuracy and efficiency of the system. The human-centered design mechanism is employed during the preprocessing stage to maintain a high recognition rate in challenging environments. Visual vocabulary recognition is performed to advance the matching stage from feature-level to object-level for faster object matching. More detailed illustrations of the two functions will be presented in the following subsections.

B. Human-Centered Design Mechanism

Compared to previous recognition platforms that focus simply on the processing speed, the proposed system aims to achieve not only efficiency but also high accuracy in challenging environments. To prevent degradation in recognition performance due to severe camera shake and dramatic changes in the object viewpoint, a human-centered design mechanism is proposed for the preprocessing stage. The mechanism consists of three stages: camera stabilization, attention tracking, and object viewpoint prediction.

The camera stabilization stage is designed to maintain high accuracy when the wearable device suffers from unwanted camera motion. As mentioned before, all frames except for the first frame of every 30 frames are processed in the coarse-grained object tracking stage. Ideally, the new ROIs for the detected target objects can be predicted according to their historical movements. For a mobile wearable device, however, dramatic global motion can be caused by severe camera shake. This unwanted camera motion results in a degradation of the tracking accuracy of the ROIs, especially the ROIs associated with moving objects. To address this problem, the main idea of the camera stabilization mechanism is to gather statistics describing the global camera motion for each frame. The information about the global camera motion is sent to the attention tracking stage. The attention tracking stage would then remove the global camera motion from the overall motion, which would leave only the local object motion from which accurate individual object movements can be captured. By statistically analyzing the displacements of the same features between two consecutive frames, the global camera motion of the mobile device can be estimated. However, the matching of corresponding features between two neighboring frames causes a huge amount of memory accessing operations and requires a large memory space. We have observed that the global camera motion is equivalent to the motion of a static object between two frames. Motivated by this, after an object with the same displacement as the global camera motion is found, this object is selected and referred to as the "static object." In this way, it is only necessary to calculate the motion displacements of the static object, and the global camera motion can be obtained. This avoids large amounts of matching of corresponding features between neighboring frames.

The attention tracking stage aims to trace the ROI for each detected object over the subsequent 29 frames. The detection strategy in the coarse-grained object tracking stage greatly reduces the computational requirements for feature extraction and matching from processing of the entire frame to processing of only a few ROIs of the detected objects. In the coarse-grained object tracking stage, the attention tracking step predicts the new position of each ROI according to its historical movements. An

accurate historical movement can be calculated by removing the global camera motion from the combined global and local motions. To accurately detect the exact position of the new ROI in the next frame, a search region around the predicted center of the new ROI is created. The attention tracking mechanism continuously records the movements of each ROI during the coarse-grained object tracking stage.

The object viewpoint prediction mechanism is employed to provide a 160° viewpoint invariance with respect to the orientation of the object. For wearable vision applications, the camera may change its orientation due to the user turning their head. In addition, the direction of movement or speed of detected target objects may also change. The above two cases would result in a dramatic change to the object viewpoint, which is a principal factor that causes a degradation in the recognition performance of the entire system. The main idea of the proposed object viewpoint prediction mechanism is to estimate the possible object poses in the next frame and synthesize images of these object poses in advance to maintain a high feature matching success rate. Based on information about the global camera motion and individual object motions obtained during the previous stages, the object viewpoint prediction mechanism estimates viewpoint parameters and predicts the possible poses of each object. In our work, a database containing 150 object images is utilized. The 150 images contain images of 50 objects from three different viewpoints (front on, 80° to the left, 80° to the right). For each frame, the object viewpoint prediction mechanism synthesizes five images, which contain the most likely images of the object pose that will appear next. For example, if the possible object viewpoint is estimated as right 30°, then the synthesized viewpoints of the object poses will be right 10, 20, 30, 40, 50°, which cover a 50° range of viewpoints. The viewpoint range can be adaptively altered from 50° and to maximum of 160° (from left 80° to right 80°) depending on the viewpoint change rate of the object across consecutive frames. Then, after the object viewpoint prediction step, the SIFT feature extraction is performed. Fig. 3 presents the recognition rate improved by the object viewpoint prediction mechanism under various object viewpoint changes. In this paper, the recognition rate is defined as the number of correctly detected ROIs with respect to the number of golden ROIs that should exist in the image. From Fig. 3, we can see that a recognition accuracy greater than 94% is achieved even when the viewpoint of the object pose changes dramatically to 80°.

In order to measure the recognition rate improved by the use of high resolution video input and the human-centered design mechanism, we compared our methods with other approaches. Fig. 4 presents the recall-precision curve for detecting an object at 30 m with an 40° change in the object viewpoint. Our proposed approach, denoted as "full HD + HCD", adds the human-centered design mechanism as the preprocessing stage of the entire system for full HD resolution input. As to the method "full HD", it only performs full-frame object recognition for each frame of the full HD resolution video without utilizing the human-centered design mechanism. The approach "VGA" only processes full-frame recognition for each frame in the VGA resolution input. The recall has the same definition

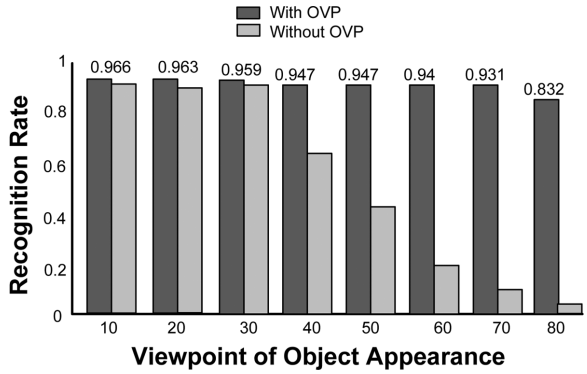

Fig. 3. Recognition rate with respect to object viewpoint changes.

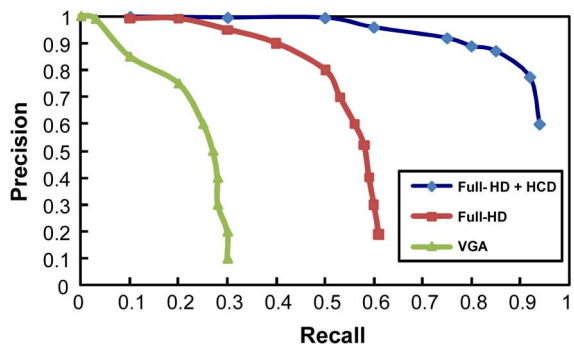

Fig. 4. Recall-precision curve.

with the recognition rate utilized in this work. The precision is the rate of the number of correctly detected ROIs with respect to the number of detected ROIs in the video. The proposed method outperforms the others, resulting in the highest absolute recall and the highest precision at virtually all values of recall. On the contrary, "full HD" has degradation of accuracy owning to the large change in the object viewpoint. "VGA" is tested with the worst recognition performance because the low resolution video is unable to provide enough features for detecting distant or small-sized objects.

C. Visual Vocabulary Recognition

Visual vocabulary recognition is employed to accelerate the speed of the matching stage, which is the performance bottleneck of most object recognition systems. The main idea of the visual vocabulary recognition approach is to advance the matching stage from feature-level to object-level, which greatly reduces the memory bandwidth required for the matching operations. In conventional object matching methods, it is intuitive to recognize an object by comparing the similarity of features extracted from the input video frame with features stored in a database. For a full HD resolution video sequence, thousands of features are extracted from one input frame. To search for the nearest neighbor of the input features in a database that stores 91 K features, nearly millions of feature comparisons have to be executed for one input feature. Such a high memory access rate severely degrades system performance and causes high power consumption. To cope with this problem, the visual vocabulary

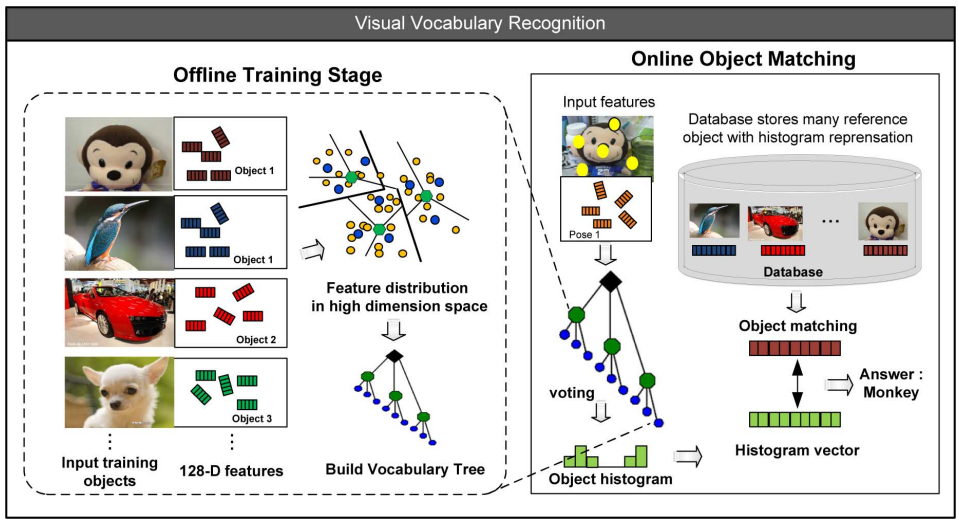

Fig. 5. Visual vocabulary recognition.

recognition mechanism, which is an object-level matching approach, is adopted in the proposed SoC. The coarse-grained object-level matching is performed based on a BoW object representation with a visual vocabulary tree.

Fig. 5 depicts the algorithm of the visual vocabulary recognition. In the training stage, the visual vocabulary tree is built to quantize the 128-D feature space by applying hierarchical K-Means clustering [17] algorithm. An example of a 3-level vocabulary tree is shown in Fig. 5 that illustrates the operations involved in visual vocabulary recognition. In this example, there are three divisions in each level, which means that each inner node has three child nodes. In the first level, all the features of objects stored in the database are classified into three clusters according to their positions in feature space. The centroid of each cluster then represents a visual word of the vocabulary tree and is described as a 128-D vector. This clustering process runs recursively until nine visual words are generated at the leaf level. Next, each feature of an object stored in the database is processed by comparing the Euclidean distance between the feature and the visual words in each level. Finally, each input feature of an object is matched to a visual word in the leaf level of the tree and votes to the bin associated with this visual word in the object histogram. By repeating this process, the voting results for an object form a 9-D histogram that is a new presentation of the object.

In the online object matching stage, all of the features extracted from the ROI of the detected object are compared to the nodes of the vocabulary tree. The generated histogram vector for each ROI is compared with histogram vectors stored in the database. Therefore, in the object matching stage, a histogram comparison for an object is only executed once instead of matching a huge number of features associated with one object. Fig. 6 depicts accuracy with respect to the number of levels of the binary tree. It can be observed that as the number of levels for a tree becomes higher than five, the recognition rate increases to more than 90%. Based on the tradeoff between the accuracy and hardware costs, a 6-level binary tree is implemented as the visual vocabulary tree in this work.

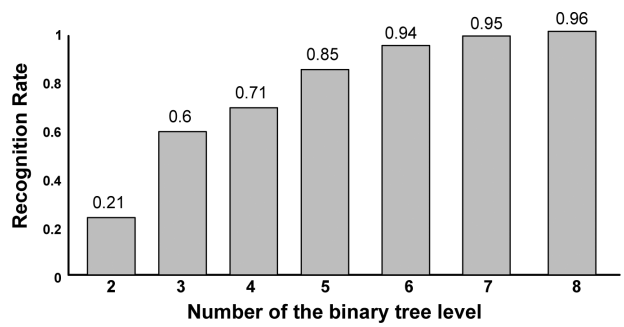

Fig. 6. Recognition rate with respect to increasing number of the binary tree level.

III. System Architecture

Fig. 7 shows the overall block diagram of the proposed SoC. It is composed of a human-centered design (HCD) engine, feature extraction (FE) engine, visual vocabulary processor (VVP), and a feature matching processor (FMP). Based on the proposed system flow illustrated in Section II, the system operations can be divided into two parts: fine-grained object recognition (once every 30 frames) and coarse-grained object tracking (the subsequent 29 frames). In the fine-grained object recognition stage, the FE and FMP are performed to recognize the target objects. The FE is responsible for SIFT feature extraction. It contains two sub-modules: the key-point detection (KD) and feature description (FD) modules. Each pixel of the input frame is processed by the KD in order to search for scale-space extrema. After that, key-points are detected and sent to the FD to form SIFT features. The FMP performs conventional feature-level matching once every 30 frames. More detailed descriptions of the FE and FMP are given in the next subsection.

For the subsequent 29 frames handled during the coarse-grained object tracking stage, the HCD, FE, and VVP are executed to provide accurate and efficient object tracking. The HCD is responsible for the human-centered design stage. The HCD is the preprocessing stage of the system, and is composed of the camera motion stabilization (CMS),

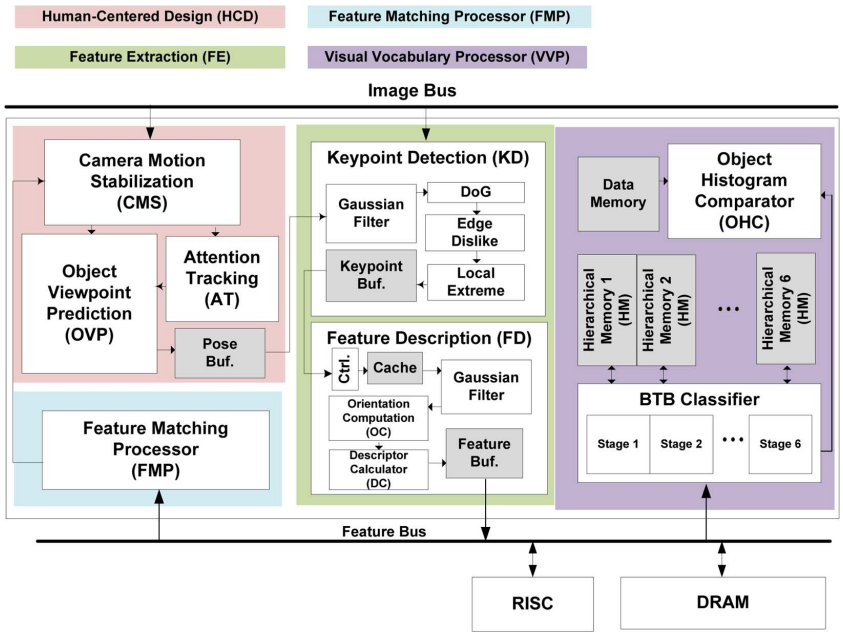

Fig. 7. System architecture.

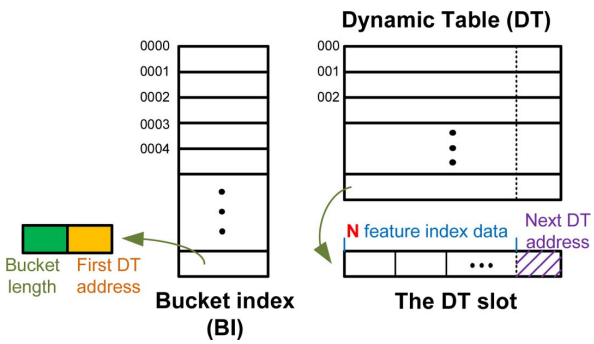

Fig. 8. Proposed hash structure based on the bucket index (BI) and dynamic table (DT) allocation.

attention tracking (AT), and object viewpoint prediction (OVP) modules. The VVP is utilized to accelerate the object-level matching stage. In the proposed system, a dual-bus structure is employed for efficient data transmission. An image bus is employed for transmitting image data from the external memory. A feature bus is used to store or fetch features from the database. The whole system runs visual recognition processing for 1920×1080 resolution video input at 30 fps.

A. Feature Extraction (FE)

The FE engine realizes 128-D SIFT feature generation. It consists of the key-point detection (KD) module and the feature description (FD) module. To achieve high efficiency, both sub-modules are implemented with dedicated hardware for real-time SIFT feature generation. The KD module is responsible for the key-point detection stage in the SIFT algorithm. It implements functionalities such as the 2-D Gaussian filter, difference of Gaussian, edge checks, and local maxima checks. In our system, three octaves and six scales of scale space are adopted in the KD to achieve good scale invariance. The KD contains three key-point processors and each processor has 18 Gaussian filters to speed up the 2-D Gaussian operations. The FD contains two sub-modules: the orientation computation (OC) module and the descriptor calculation (DC) engine. The OC first computes the gradient orientation and magnitude for each pixel located in the 16×16 region around the detected key-point. Next, for each region, the OC calculates the major orientation by constructing a 36-bin orientation histogram and adding each pixel in the search region to a histogram bin according to its weighted gradient magnitude. The computation results of orientations and magnitudes for pixels in the region are also sent to the DC for 128-D SIFT descriptor generation. Within the 16×16 region around the key-point, 16 8-bin orientation histograms are constructed to store the orientations and magnitudes of pixels located in each 4×4 sub-region.

However, the calculation of the magnitude and orientation for a pixel require the arctangent and square root operations that are complex and would significantly increase the hardware costs of the OC. In order to reduce the area of the calculations, we utilize two lookup tables that contain the trigonometric and square root values to alleviate the complex computations. In this way, the processing time of arctangent and square root calculations are accelerated by three times while the area of the OC is reduced by 11% compared to a general arithmetic logic unit (ALU). In addition, the DC has the capability to compute two features in parallel with sharing arithmetic calculations such as the arctangent, division, square root, and sine operations. Overall, the processing time for generating one feature by the FE is 110.9 us at 200 MHz operating frequency. Due to the repeated multiplication and division in the FE, it generally requires high power consumption. In this work, we try to reduce the power consumption of the FE and accelerate the processing of recognition by performing the FE only once every 30 frames. The other 29 frames are classified in the coarse-grained object tracking stage where only ROIs containing the information of the target objects need to be processed by the FE. For the case of tracking five ROIs

in parallel, the power consumption of the FE can be reduced by 77.3% compared to the full-frame feature extraction.

B. Feature Matching Processor

The FMP is responsible for feature-level matching in the fine-grained object recognition stage. Feature-level matching is performed once every 30 frames. For each feature extracted from the first frame of every 30 frames, the FMP searches through all the features stored in the database for its nearest neighbor in Euclidean space. However, the computation of the nearest neighbor searching is very time-consuming in a database containing millions of features. According to the software simulation result for a brute force searching, it costs around 0.7 s for an input feature to find the nearest neighbor in a database that contains 91 K features. To accelerate the feature matching stage, tree-structured algorithms such as the KD-tree [18] are widely adopted to speed up the feature matching stage in traditional recognition approaches. Unfortunately, tree-structured approaches do not show significant performance improvements compared to brute force searching when the dimensions of the feature descriptor are more than 20. Reference [19] indicates that hash-based searching can process high-dimensional features more efficiently than tree-structured searching. Motivated by this, we design a locality-sensitive hashing (LSH) [20] based FMP to reduce the searching time for 128-D SIFT feature matching. The basic idea behind the LSH algorithm is to hash the input features by considering their spatial relationships in Euclidean space so that there is a high probability that similar features will be mapped to the same bucket or to a neighboring bucket. In this way, feature-level matching is simplified to a search through the hash table for a small set of features in just a few buckets. Compared with nearly 91 K comparisons for one input feature to find its nearest neighbor using the brute force method and the searching approach using the KD-tree, the comparisons for one feature in the LSH algorithm can be significantly reduced to only 301 times on average according to our experiment result.

Despite of the efficiency of the hash-based searching method, a large amount of memory is required to construct the hash tables. This is because a large memory space for the bucket for each hash index needs to be allocated in advance just in case a large number of features are mapped to the same bucket. From our experiments, we have found that a mere 10% of the memory space allocated for the hash tables is utilized. In this paper, a new dynamic hash table allocation mechanism is proposed to reduce this unnecessary memory allocation. Fig. 8 depicts the proposed hash structure, which is composed of the bucket index (BI) and the dynamic table (DT). The BI consists of a 6-bit bucket length that indicates the number of features stored in the corresponding bucket in the DT, and a 10-bit address, which reveals the starting address of this bucket in the DT. A bucket is composed of a linked list of slots. In the DT, each slot has N feature indexes and one address pointer that links to the next slot in the DT.

Fig. 9 illustrates the detailed architecture of the proposed hash-based searching mechanism. In the proposed system, three hash tables are constructed to increase the versatility of the hash-based searching mechanism. More hash tables would increase the matching accuracy by involving different sets of feature

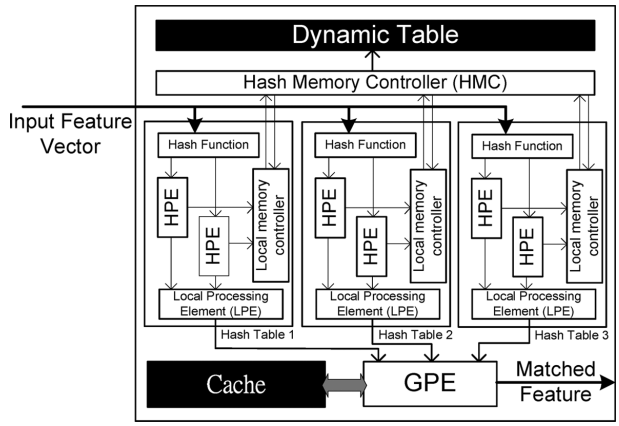

Fig. 9. Proposed feature matching diagram.

to be matched. The hash function (HF) engine is adopted to map each feature stored in the database to the corresponding bucket in the DT. The hash processing element (HPE) helps to check whether there is enough space in the current slot when inserting new data. The hash memory controller (HMC) assists in memory allocation and management of the DT. Through dynamically allocating the bucket memory, the table utilization rate is increased to more than 60%, which is six times larger than the original hash table structure. For each DT, two HPE are utilized in order to process two features in parallel. When feature matching is performed, the local processing element (LPE) is responsible for searching in the database for the nearest neighbor of a feature by collecting similar features from the DT and sending the results to the global processing element (GPE). The GPE removes identical features collected from different hash tables to avoid redundant comparisons. In addition, in order to reduce redundant fetching of the same feature from the external memory, a 64-address two-way set-associative cache is adopted in the FMP. A 52% reduction of redundant feature fetching is achieved due to the cache mechanism.

C. Human-Centered Design Module

Fig. 10 depicts the detailed architecture of the HCD. The HCD consists of the CMS, AT, and OVP modules. The HCD aims to assist the proposed system in achieving a high recognition rate even when faced with a challenging environment, such as that caused by severe camera shake or dramatic object viewpoint changes. To the best of our knowledge, state-of-the-art recognition systems have not been designed to cope with these real challenges. The HCD is the first proposed preprocessing module dedicated to a wearable visual recognition system.

As mentioned in Section II-A, to overcome the camera shake problem, the basic design concept of the CMS is to maintain the tracking accuracy by gathering statistics describing the global movement of the camera between frames. In this way, the AT can utilize the global motion information to accurately extract the motion of individual objects. In order to achieve this goal, the CMS calculates the global camera motion for each frame by computing the displacements of features that have been matched in two consecutive frames. We refer to the displacement of a feature between two frames as the feature motion vector. The CMS uses a statistical method to estimate the orientation and

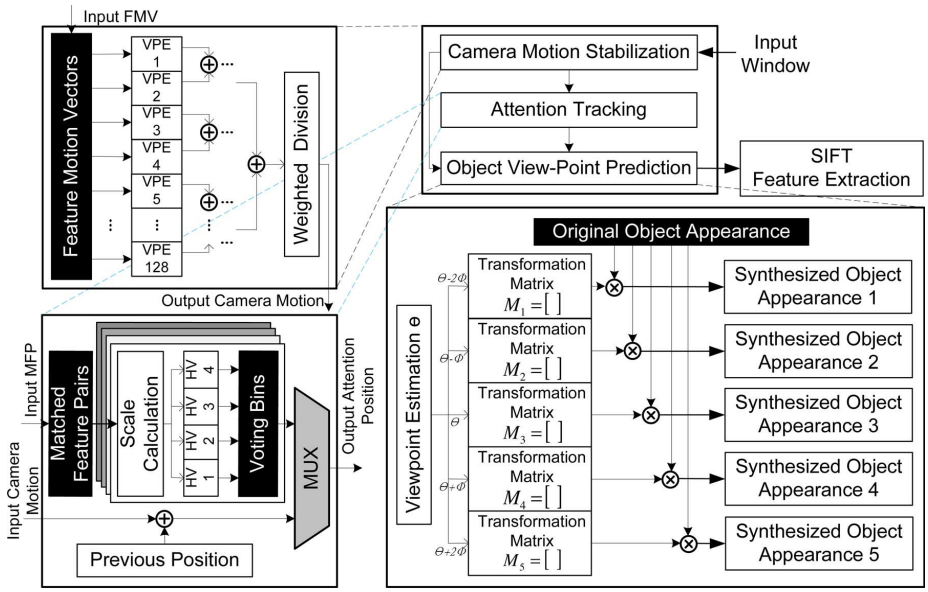

Fig. 10. Architecture of the proposed human-centered design.

magnitude of the global camera motion between two consecutive frames from all the corresponding feature motion vectors. Note that as the camera moves continuously, new features are added and old features are removed from the list of existing features that appear in the video stream. Previous studies [21] indicated that a better estimate of the global motion can be obtained by analyzing features that have been continuously detected in the video stream for a reasonably period of time. Thus, the confidence index is adopted to represent the weighting of a feature. The longer the feature exists in the video stream, the higher the value of its confidence index will be. To implement this concept, every feature is assigned a unique 4-bit confidence weighting index. In the CMS, 128 parallel vector processing elements (VPE) are designed to update the value of the confidence index for each feature and compute its weighted feature motions. The following is a seven-level accumulator handles the weighting of the feature motions used for the global motion calculation. In this way, a maximum of 128 features can be analyzed concurrently to determine the global camera motion. Immediate results for the 128 VPEs can be obtained after eight cycles. The processing of the global camera motion runs in a pipeline manner and requires only 16 cycles when 1024 features are analyzed.

The AT is employed to generate the ROIs of the target objects. In the fine-grained object recognition stage, the AT executes the GHT algorithm to initialize the ROI for each recognized object. During this stage, the basic concept of the AT is to cluster the matched features that correspond to the same object poses for each video frame. Each matched feature in the video frame has to vote to one pose of an object that is stored in the database. The object pose is determined on the basis of the differences between the scale, rotation, and x and y coordinates of the two matching features. These four parameters contain "4-D invariance information." To execute the complex voting mechanism of the GHT algorithm, eight parallel processors are used and each processor has one scale calculator (SC) and four hough-table voting (HV) processors. Firstly, the SC calculates the scale difference between the two matching features. Eight matching pairs can be processed in parallel. Next, the HV is responsible for the voting process for each matching pair of features according to their 4-D invariance information. During the coarse-grained object tracking stage, the AT is responsible for computing the individual object movements and predicting the possible positions of the ROIs in the next frame. A predicted accumulator that estimates the possible positions of new ROIs by analyzing the previous ROI movements and the global camera motion is implemented in the AT.

To overcome the difficulties caused by dramatic object viewpoint changes, the main idea of the OVP is to predict the possible object pose that may appear in the next frame and provide multiple potential poses of the objects in order to enhance the matching performance. In the OVP, the viewpoint estimation engine is firstly executed to generate the viewpoint parameters. These parameters are inferred from information about the global camera motion and the individual object motions obtained during previous stages. Next, five parallel transformation matrix generators concurrently synthesize five object poses with multiple viewpoints. Each transformation matrix generator is designed as a SIMD processing element to exploit pixel-level parallelism. The maximal range of the synthesized viewpoint is 160° (80° for each side). The following stage is the FE performed for SIFT feature extraction.

D. Visual Vocabulary Processor

To accelerate the speed of object matching to enable real-time processing, a highly parallel architecture, i.e., the VVP, is employed in the SoC. The basic design concept of the VVP is to use BoW object-level matching and utilize the hierarchical memory and parallel structures to reduce the large memory accessing requirements during the matching stage. The detailed architecture is shown in Fig. 11. The VVP contains three parts: the hierarchical memory (HM), the binary-tree-based classifier

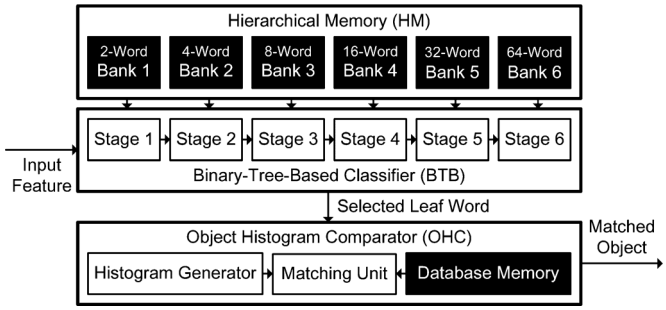

Fig. 11. VVP architecture.

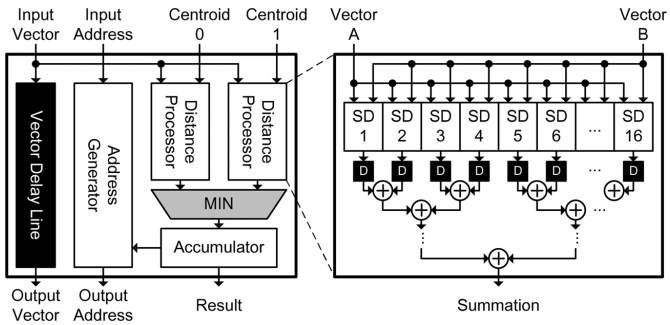

Fig. 12. Detailed architecture in each stage of BTB.

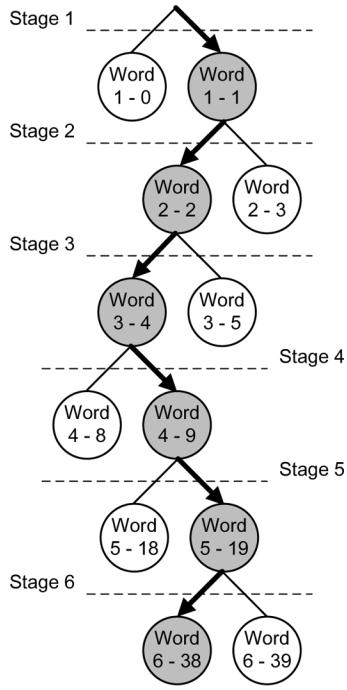

Fig. 13. Example of the word visiting process.

(BTB), and an object histogram comparator (OHC). The BTB is responsible for classifying each 128-D input feature into a leaf node in the binary tree. The OHC generates a histogram for the detected object by accumulating the classification results from the BTB. Finally, an object histogram comparison is performed for object matching. The HM stores the visual words contained in the vocabulary tree. In this work, the vocabulary tree is built as a six-level binary tree with 126 words in total. In the training phase, 91 K features are used to train the binary tree before using the VVP for object detection. The 64 leaf nodes of the binary tree divide the 128-D feature space into 64 non-overlapping partitions. The HM has six banks of memories to offer tree-node information to the BTB to enable the determination of the word that is closest to the input vector. Fig. 11 shows the six-stage architecture of the BTB, and the detailed architecture of each stage is depicted in Fig. 12. There are two distance processors in every stage of the BTB. Each of them contains 16 square-of-a-difference (SD) calculators and a tree-like adder for parallel computations of Euclidean distances. The distance processors are connected to their corresponding banks in the HM. The six-stage architecture of the BTB requires only eight cycles to process a 128-D feature with a 16 dimensions/cycle throughput. Compared with the conventional feature-level matching method that costs bandwidth of 4.9 GB/s, the VVP only consumes 154 MB/s on average, achieving 97% of memory bandwidth reduction.

An example of the word visiting process in the six-level binary tree is shown in Fig. 13. First, the SD calculators in the first stage of the BTB compute the Euclidean distance between the input feature and Word 1–0 and the Euclidean distance between the input feature and Word 1–1. Both Word 1–0 and Word 1–1 are stored in the first bank of the HM. The result in Fig. 13 shows that Word 1–1 is closer in feature space to the input feature than Word 1–0. The input feature is then sent to the second stage of the BTB. Similarly, the SD calculators in the second stage of the BTB compute the distance between the input feature and the corresponding words: Word 2–2 and Word 2–3. Because Word 2–2 is nearer to the input feature than Word 2–3, the input feature is sent to the third stage of the BTB, and the distances between it and the corresponding words are computed by accessing the HM. The same procedure repeats until six stages are complete, and the input feature is finally classified into Word 6–38, which is the leaf level of the binary tree. Next, the OHC records the classified result by increasing the value of the bin in the object histogram that corresponds to Word 6–38. After processing all the features in the ROI of the detected object, the generated object histogram vector is compared with all of the object histograms stored in the database. Based on this BoW object matching concept, the whole matching process can be handled at the object level. Instead of matching thousands of features of two objects in traditional matching methods, only one operation of matching histograms is required when comparing two objects with the BoW method.

IV. Chip Implementation and Evaluation

A. Chip Implementation

A prototype chip for the proposed recognition SoC is fabricated by TSMC with a 65 nm 1P9M process. Table I lists the detailed chip features and specifications, and a chip micrograph is shown in Fig. 14. The chip size is 6.38 mm^2 including the IO and bonding pad. The total gate count is 0.91 M and the total on-chip SRAM memory is 40 kB. The maximum operating frequency of the proposed system is 200 MHz and the peak performance achieves 164.95 GOPS. The average power consumption of the system is 52 mW at the supply voltage of 1.0 V while running at 200 MHz. The peak power consumption is 198.4 mW. The chip

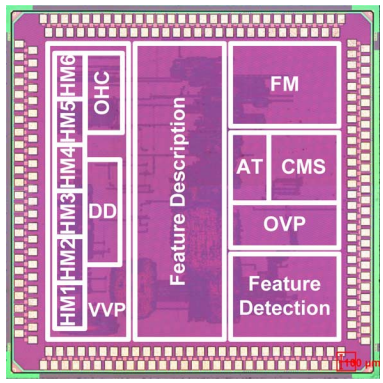

Fig. 14. Chip photograph.

TABLE I
CHIP SUMMARY

Item	Specification
Die Size	2.5mm × 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(1920×1080 30fps) Video

supports multi-object recognition for videos with 1920 × 1080 resolution at 30 fps.

B. Chip Comparison

Table II summarizes the comparison between the proposed system and previous recognition systems [6], [10]. From an applied perspective, the previous works did not address the issues of degraded accuracy resulting from frequent camera motions and dramatic object viewpoint changes. Our system aims to solve these problems, which are often present in the real world. Our method supports 160° object viewpoint invariance, and maintains high recognition accuracy even in challenging environments.

From an efficiency perspective, in [6], the processing time during the feature matching stage is at least 3 M cycles for 256 features. This processing speed of the matching stage is unable to handle real-time full HD video processing, where at least thousands of features need to be extracted. In our system, the proposed VVP greatly contributes to speeding up the processing time of the matching stage by advancing matching from the feature-level to the object-level. The VVP performs object matching with only one histogram comparison that costs nine cycles (eight cycles to form the object histogram and one cycle for the object comparison). As a result, our system has the ability to process full HD videos at 30 fps, compared to previous works, which could only handle real-time VGA resolution videos. As mentioned in Section I, compared to VGA recognition processing, full HD detection can greatly enhance

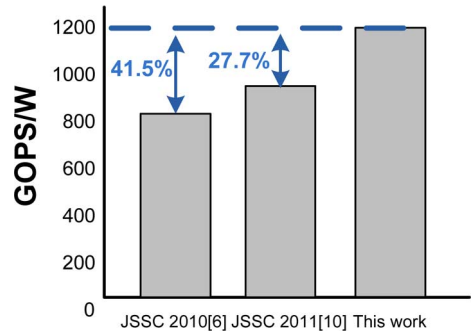

Fig. 15. Power efficiency comparison.

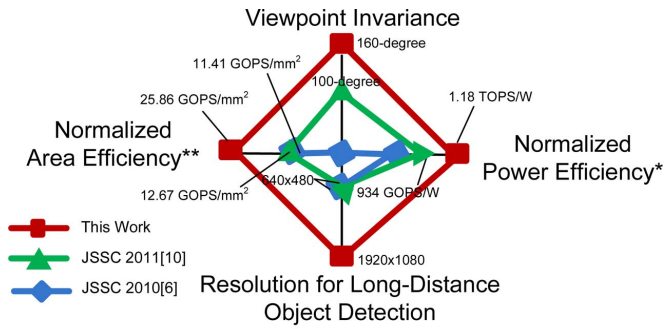

Fig. 16. Overall system comparison with previous works.

the recognition accuracy when detecting distant objects, which is a basic requirement for outdoor recognition activities.

From a power consumption perspective, our system outperforms previous works. The comparison for power efficiency is shown in Fig. 15. Note that the technology is scaled from a 0.13 m process to a 65 nm process. Based on the adopted object-level matching via the VVP, a 97% reduction of memory bandwidth in the matching stage is achieved. Thus, the average power efficiency of the proposed work reaches 1.18 TOPS/W, which is much better than that reported in previous works. From a cost perspective, our system achieves an area efficiency of 25.9 GOPS/mm^2, which is also superior to previous works. Fig. 16 presents an overall system comparison. As shown in this figure, the proposed system outperforms the existing systems in both functionality and efficiency.

C. Evaluation

For chip-level verification, the Agilent 93000 mixed-signal SoC test system is used for chip testing. The maximum measured frequency is 200 MHz, which is sufficient for processing the specified targets. In our work, eight 1920 × 1080 videos are tested to evaluate the robustness of the proposed system. These testing sequences are recorded from a video camera which was mounted on a user's head to simulate the real user experiences, including navigating in various indoor and outdoor places. Four videos are recorded in outdoor environments, which are referred to as "street," "road," "campus," and "corridor." The other videos are recorded in indoor environments. These indoor environments are referred to as "supermarket", "laboratory", "living room," and "library." Fig. 17 shows the testing system,

TABLE II
COMPARISON WITH PREVIOUS WORKS

	JSSC 2010 [6]	JSSC 2011 [10]	This Work
Object Viewpoint Functionality	Not Supported	100 degrees	160 degrees
Resolution	VGA(640×480)	VGA(640×480)	full HD(1920×1080)
Average Power Consumption	496mW	345mW	52mW
Peak Power Consumption	695mW	704mW	198.4mW
Technology	0.13μm	0.13μm	65nm
Logic Gate Count	3.73M	2.92M	0.91M
On-Chip SRAM	396KB	612KB	40KB
Die Size	7mm×7mm	10mm×5mm	2.5mm×2.6mm

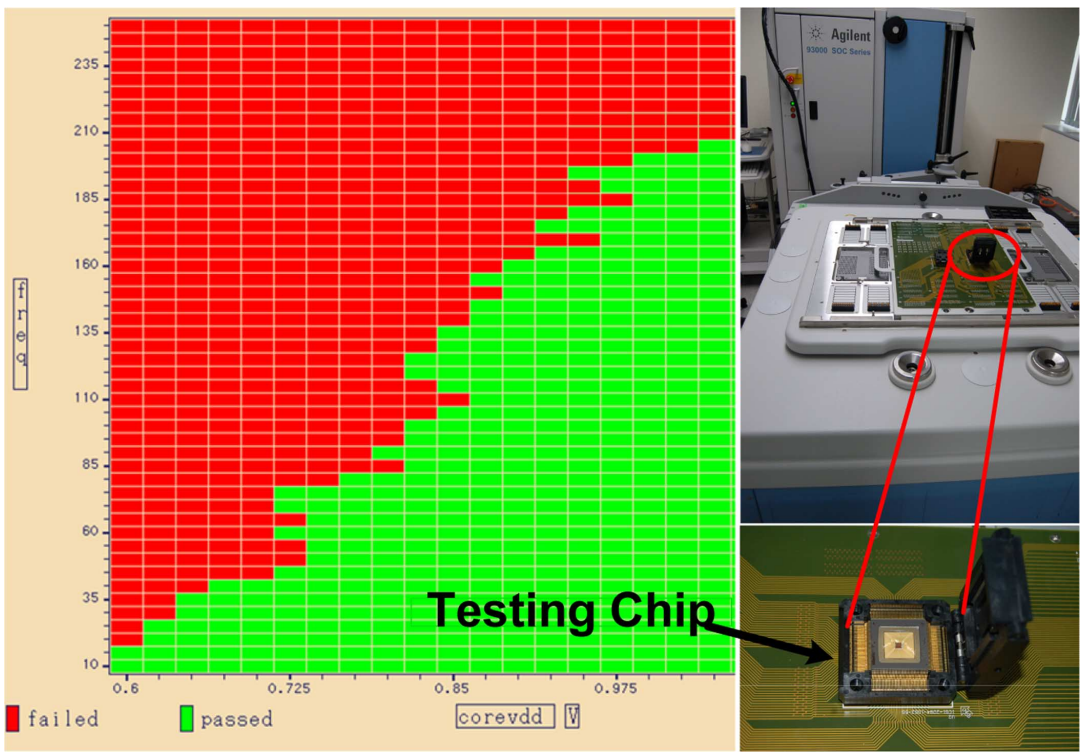

Fig. 17. SHMOO plot generated by Agilent 9300 mixed-signal SoC test system.

fabricated chip and the SHMOO plot. In general, multi-object tracking costs more computational power due to more features that need to be extracted. Since the ROI size of an object may change according to the distance between the object and the user, the number of features extracted from the object ROI is not fixed in a testing sequence, thereby we firstly tested the maximal number of object ROIs that the system can process in parallel. According to the evaluation results, our system can support up to five object tracking in real-time performance. In outdoor environments, the target objects are basic things such as traffic lights, cars, and logo signs. As to indoor environments, soda bottles, magazines, furniture, cups and computer peripherals that are present in the videos are chosen as the target objects.

First, we tested the reliability of the chip in terms of the recognition rate. For each video, three object ROIs that are present in 90% of video frames in the sequence are set as the target objects. The average recognition rate is 94%. According to the testing results, the accuracy depends on the scene of the video input. Among the eight sequences, the video "street" achieved the best recognition rate (98%). This is because the traffic lights and logo signs have clear corners, which are features that are easily discriminated and extracted by the system. In addition, the scene background of the video "street" is not complex. The video "supermarket" resulted in the lowest recognition rate (89%). This is because the targets objects, which were several specific brands of soda bottles with textural logos attached, appear beside many other brands of bottles that are also tagged with landmarks or logos of a complex texture. When dealing with videos containing textural content, especially in situations where many similar things are placed together, more features are extracted and these features are close to each other in Euclidean space. This situation would result in the degradation of the recognition performance.

Second, we tested the relationship between the recognition performance and the number of object ROIs. For each sequence,

three ROIs and five ROIs of object tracking are tested separately. According to the evaluation results, we cannot see that the number of ROI has significant impact on the recognition accuracy. The difference of the average recognition rate between tracking for three ROIs and five ROIs is less than 0.9%. This is because at least hundreds of features can be extracted from the ROI in a full HD resolution image rather than only few features could be found in the ROI of an image with low resolution. This ensures the stability of the object histograms of the target objects generated by the VVP. As a result, the matching performance is still reliable with respect to the increased number of ROIs. Meanwhile, in the fine-grained object recognition stage, the proposed system executes the GHT algorithm that can help to filter true matching pairs of features. In this way, the proposed system can achieve in both high recognition rate and precision. Lastly, we observed that the scene contains complex background or textural content also resulted in higher power consumption. The video "supermarket" achieved the maximal power of 74.88 mW among the eight sequences when tracking five ROIs in parallel at 200 MHz. This is because 95% of the content of the video are beverage bottles attached with textural logos. The textural content and complex background that increase on average 21% of the number of extracted features for one frame in the video than those of other videos.

V. Conclusion

In this paper, a 1920×1080 160° object viewpoint recognition SoC is proposed to achieve highly accurate object recognition with low power consumption that is suitable for wearable vision applications. Our contribution can be summarized as follows. First, our system supports full HD resolution video at 30 fps for long-distance object recognition in outdoor environments. The accuracy of the detection of an object 50 m away is greatly enhanced from the 29% that is achievable with VGA resolution video input to 95% by using full HD resolution video sequences. Second, the HCD mechanism consisting of CMS and OVP modules provides reliability and stability for the proposed system in challenging environments. To avoid degrading the recognition rate during periods of severe camera shake, the CMS estimates the global camera motion to maintain accurate object tracking. With the prediction of multiple object viewpoints, the OVP achieves a recognition accuracy of 94% on average even during dramatic changes in the object viewpoint. Last, a VVP is employed to advance the matching stage from feature-level to object-level. The adopted object-level BoW matching and the massively parallel architecture of the VVP contribute to a 97% reduction in the memory bandwidth required for the matching stage. This innovation solves the bottleneck problem of traditional object recognition systems. The visual recognition SoC is realized on a 6.38 mm^2 die with 65 nm CMOS technology. Only 52 mW of power is dissipated and a power efficiency of 1.18 TOPS/W and an area efficiency of 25.9 GOPS/W are achieved. Validation results demonstrate that our system achieves high accuracy and efficiency. To summarize, the high tolerance of the system to camera shake and changes in object viewpoints provided from the HCD and the low power consumption achieved by the VVP make the proposed system suitable for wearable visual applications.

Acknowledgment

The authors would like to thank TSMC University Shuttle Program and Jefferson Hsieh for process support. They would also like to thank Chip Implementation Center (CIC) for design flow supporting and chip testing.

References

[1] D. G. Lowe, "Distinctive image features form scale-invariant keypoints," *Intl. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[2] Y. Ke and R. Sulkthankar, "PCA-SIFT: A more distinctive representation for local image descriptors," in *Proc. IEEE CVPR*, 2004, pp. 506–513.
[3] K. Kim, K. Kim, J. Y. Kim, S. Lee, and H. J. Yoo, "An 81.6 GOPS object recognition processor based on NoC and visual image processing memory," in *Proc. IEEE CICC*, 2007, pp. 443–446.
[4] Abbo, R. Kleihorst, V. Choudhary, L. Sevat, P. Wielage, S. Mouy, and M. Heijligers, "XETAL-II: A 107 GOPS, 600 mW massively-parallel processor for video scene analysis," *IEEE J. Solid-State Circuits*, vol. 43, no. 1, pp. 192–201, Jan. 2008.
[5] K. Kim, S. Lee, J. Y. Kim, M. Kim, and H. J. Yoo, "A 125 GOPS 583 mW network-on-chip based parallel processor with bio-inspired visual attention engine," *IEEE J. Solid-State Circuits*, vol. 44, no. 1, pp. 136–147, Jan. 2009.
[6] J.-Y. Kim, M. Kim, S. Lee, J. Oh, K. Kim, and H. J. Yoo, "A 201.4 GOPS 496 mW real-time multi-object recognition processor with bio-inspired neural perception engine," *IEEE J. Solid-State Circuits*, vol. 45, no. 1, pp. 32–45, Jan. 2010.
[7] M. Hiromoto, H. Sugano, and R. Miyamoto, "Partially parallel architecture for Adaboost-based detection with Haar-like features," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 1, pp. 41–52, Jan. 2009.
[8] V. Bonato, E. Marques, and G. A. Constantinides, "A parallel hardware architecture for scale and rotation invariant feature detection," *IEEE Trans. Circuits Syst.*, vol. 18, no. 12, pp. 1703–1712, Dec. 2008.
[9] K. Mizuno, H. Noguchi, G. He, Y. Terachi, T. Kamino, H. Kawaguchi, and M. Yoshimoto, "Fast and low-memory-bandwidth architecture of SIFT descriptor generation with scalability on speed and accuracy for VGA video," in *Proc. IEEE FPL*, 2010, pp. 608–611.
[10] S. Lee, J. Oh, J. Park, J. Kwon, M. Kim, and H. J. Yoo, "A 345 mW heterogeneous many-core processor with an intelligent inference engine for robust object recognition," *IEEE J. Solid-State Circuits*, vol. 46, no. 1, pp. 42–51, Jan. 2011.
[11] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 10, no. 27, pp. 1615–1630, 2005.
[12] Y. C. Su, K. Y. Huang, T. W. Chen, Y. M. Tsai, S. Y. Chien, and L. G. Chen, "A 52 mW full HD 160-degree object viewpoint recognition SoC with visual vocabulary processor for wearable vision applications," in *Symp. VLSI Circuits Dig. Tech. Papers*, pp. 258–259.
[13] G. Csuka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. IEEE ECCV Workshop Statistical Learning CV*, 2004, pp. 59–74.
[14] J. Yang, Y. G. Jiang, A. Hauptmann, and C. W. Ngo, "Evaluating bag-of-words representations in scene classification," in *Proc. ACM MIR*, 2007, pp. 197–206.
[15] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE CVPR*, 2006, pp. 2161–2168.
[16] D. H. Ballard, "Generalizing the Hough transform to detect arbitrary shapes," *Pattern Recognit.*, vol. 13, no. 2, pp. 111–122, 1981.
[17] T. W. Chen and S. Y. Chien, "Flexible hardware architecture of hierarchical K-means clustering for large cluster number," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 19, no. 8, pp. 1336–1345, Aug. 2011.
[18] S. Beis and D. G. Lowe, "Shape indexing using approximate nearest neighboring search in high-dimensional spaces," in *Proc. IEEE CVPR*, 1997, pp. 1000–1006.
[19] A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via hashing," in *Proc. VLDB*, 1999, pp. 518–529.
[20] A. Andoni and P. Indyk, "Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions," in *Proc. IEEE FOCS*, 2006, pp. 459–468.
[21] K. Y. Lee, Y. Y. Chuang, B. Y. Chen, and M. Ouhyoung, "Video stabilization using robust feature trajectories," in *Proc. IEEE ICCV*, 2009, pp. 1379–1404.

Yu-Chi Su (S'10) received the M.S. degree from the Department of Computer Science, National Tsing Hua University (NTHU), Taipei, Taiwan, in 2005. She is currently pursuing the Ph.D. degree from the DSP/IC Laboratory, Graduate Institute of Electrical Engineering, National Taiwan University (NTU), Taipei, Taiwan.

Her current research interests include intelligent visual recognition, semantic video analysis, machine learning, and human-computer interaction. She also majors in VLSI architecture design and system-on-chip design.

Keng-Yen Huang (S'10) received the B.S. degree from the Department of Electrical Engineering and the M.S. degree from the Graduate Institute of Electronics Engineering (GIEE), National Taiwan University (NTU), Taipei, Taiwan, in 2009 and 2011, respectively.

He is currently with MediaTek, Hsinchu, Taiwan. His research interests include computer vision and associated VLSI architectures design.

Tse-Wei Chen (S'07–M'11) received the B.S. degree from the Department of Electrical Engineering, National Taiwan University (NTU), Taipei, Taiwan, in 2006, and the Ph.D. degree from the Graduate Institute of Electronics Engineering (GIEE), NTU, in 2010.

In 2010, he was a foreign joint researcher with the Graduate School of Information Science, Nagoya University, Nagoya, Japan. He is currently a postdoctoral researcher with the VLSI Design and Education Center (VDEC), the University of Tokyo, Tokyo, Japan. His research interests include multimedia content analysis, computer vision, pattern recognition, machine learning, and associated VLSI architectures.

Dr. Chen was awarded the postdoctoral fellowship from the National Science Council (NSC), Taiwan, in 2010.

Yi-Min Tsai (S'10) received the B.S. degree from the Department of Electrical Engineering, National Taiwan University (NTU), Taipei, Taiwan, in 2007, where he is currently pursuing the Ph.D. degree from DSP/IC Laboratory, Graduate Institute of Electronic Engineering.

His current research interests include digital video signal processing, intelligent signal processing, 3-D video application, and image/object recognition, machine learning, computer vision algorithms, and compressive sensing. He also majors in VLSI architecture design, digital circuit design, and cell-based design flow.

Shao-Yi Chien (S'99–M'04) received the B.S. and Ph.D. degrees from the Department of Electrical Engineering, National Taiwan University (NTU), Taipei, Taiwan, in 1999 and 2003, respectively.

During 2003 to 2004, he was a research staff with Quanta Research Institute, Tao Yuan County, Taiwan. In 2004, he joined the Graduate Institute of Electronics Engineering and Department of Electrical Engineering, NTU, as an Assistant Professor. Since 2008, he has been an Associate Professor. His research interests include video segmentation algorithm, intelligent video coding technology, perceptual coding technology, image processing for digital still cameras and display devices, computer graphics, and the associated VLSI and processor architectures. He has published over 120 papers in these areas.

Dr. Chien serves as an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and *Springer Circuits, Systems and Signal Processing (CSSP)*. He also served as a Guest Editor for *Springer Journal of Signal Processing Systems* in 2008. He also serves on the technical program committees of several conferences, such as ISCAS, A-SSCC, and VLSI-DAT.

Liang-Gee Chen (S'84–M'86–SM'94–F'01) received the B.S., M.S., and Ph.D. degrees in electrical engineering from National Cheng Kung University, Tainan, Taiwan, in 1979, 1981, and 1986, respectively.

In 1988, he joined the Department of Electrical Engineering, National Taiwan University (NTU), Taipei, Taiwan. During 1993–1994, he was a Visiting Consultant in the DSP Research Department, AT&T Bell Labs, Murray Hill, NJ. In 1997, he was a Visiting Scholar of the Department of Electrical Engineering, University of Washington, Seattle. During 2004–2006, he was the executive Vice President and General Director of the Electronics Research ad Service Organization (ERSO) of the Industrial Technology Research Institute (ITRI), Taiwan. Since 2007, he serves as a Co-Director General of National SoC Program. Currently, he is the Deputy Dean of Office of Research and Development and the Distinguished Professor of Department of Electrical Engineering, NTU. His research interests include DSP architecture design, video processor design, and video coding systems. He has over 400 publications and 30 patents.

Dr. Chen has served as an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY in 1996–2008, as Associate Editor of the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS in 1999–2001, and as Associate Editor of the IEEE TRANSACTIONS CIRCUITS AND SYSTEMS—PART II: EXPRESS BRIEFS in 2000–2001. He has been the Associate Editor of the *Journal of Circuits, Systems, and Signal Processing (CSSP)* in 1999–2008, and a Guest Editor for the *Journal of Video Signal Processing Systems*. He has been an Associate Editor for the *Journal of Information Science and Engineering (JISE)* from 2002. Since 2007, he has served as an Associate Editor of *Research Letter in Signal Processing* and for *EURASIP Journal on Advances in Signal Processing*. During 2001–2004, he was also the Associate Editor of the PROCEEDINGS OF THE IEEE. He was the General Chairman of the 7th VLSI Design/CAD Symposium in 1995 and of the 1999 IEEE Workshop on Signal Processing Systems: Design and Implementation. He is the Past-Chair of Taipei Chapter of IEEE Circuits and Systems (CAS) Society, and is a member of IEEE CAS Technical Committee of VLSI Systems and Applications, the Technical Committee of Visual Signal Processing and Communications, and the IEEE Signal Processing Technical Committee of Design and Implementation of SP Systems. He was the Chair-Elect of the IEEE CAS Technical Committee on Multimedia Systems and Applications. During 2001–2002, he served as a Distinguished Lecturer of IEEE CAS Society. He has been the program committee member of IEEE ISSCC in 2004–2007. He will be the TPC chair of 2009 IEEE ICASSP and ISCAS 2012. He received the Best Paper Award from the R.O.C. Computer Society in 1990 and 1994. Annually from 1990 to 2005, he was a recipient of Long-Term (Acer) Paper Awards. In 1992, he was a recipient of the Best Paper Award of the 1992 Asia-Pacific Conference on circuits and systems in the VLSI design track. In 1993, he received the Annual Paper Award of Chinese Engineer Society. In 1996, 2000, and 2002, he received the Outstanding Research Award from the National Science Council, and in 2000, the Dragon Excellence Award from Acer. His students have won the DAC/ISSCC Student Design Contest four times since 2004, and won the Student Paper Contest at ICASSP 2006. He is a member of Phi Tau Phi.