A Hardware Accelerator for Bag-of Features based Visual Word Transformation in Computer Aided Diagnosis for Colorectal Endoscopic Images

Tetsushi Koide[1], Takumi Okamoto[1], Koki Sugi[1], Tatsuya Shimizu[1], Anh-Tuan Hoang[1],
Toru Tamaki[2], Bisser Raytchev[2], Kazufumi Kaneda[2],
Shigeto Yoshida[3], Hiroshi Mieno[3], and Shinji Tanaka[4]
[1] Research Institute for Nanodevice and Bio Systems (RNBS), Hiroshima University
[2] Graduate School of Engineering, Hiroshima University
[3] Department of Gastroenterology, Hiroshima General Hospital of West Japan Railway Company
[4] Department of Endoscopy and Medicine Graduate School of Biomedical and Health Science, Hiroshima University

1-4-2 Kagamiyama, Higashi-Hiroshima, 739-8527, Japan
E-mail: koide@hiroshima-u.ac.jp

Abstract: This paper presents an FPGA based hardware accelerator for feature transformation in real-time computer-aided diagnosis (CAD) system for colorectal endoscopic images with narrow-band imaging (NBI) magnification. We have demonstrated the proposed FPGA implementation, which was very compact, the fulfiled performance requirement from the clinical doctors (throuthput > *5 fps*, latency < *1 sec*). It achiveded throughput: *16.7 fps* and latency: *60 msec* without quality reduction of real-time diagnostic support.

1. Introduction

With the increase in the number of colorectal cancer patients, systems which support a doctor's diagnosis have been researched. The CAD system for colorectal endoscopic images with NBI magnification [1] has already been proposed [2]. The proposed CAD system identifies 3 types of endoscopic image (Type A, Type B, and Type C3) as shown in Fig. 1. Currently our software implementation of the system is able to identify with only the region of 120x120 pixels at 14.7 fps and it takes about 20 minutes to process a whole Full-HD (1920x1080) image. Further improvement in the speed is needed for realization of high performance Full HD image recognition. Our system performance must be satisfied with a demand on the clinical doctors, throughput is within from larger than 5 fps and the latency is at least within 1 second for on-the-fly diagnostic supporting. Therefor this paper proposes a hardware implementation of high speed feature transform for the CAD system.

2. Outline of Computer-Aided Diagnosis System

Outline of the proposed CAD system is shown in Fig. 2. The system is based on a Bag-of-Features (BoF) representation of local features in the endoscopy image. In feature extraction, an input image is processed as Scan Window (SW). The local feature quantities are extracted at all key points, at which the feature extraction performed in the SW. We use the Dense Scale-Invariant Feature Transform (D-SIFT) that takes key points to dense. First, the features obtained from the images of each type in learning phase are clustered based on Dense Scale-Invariant Feature Transform (D-SIFT) algorithm [3], and the center of each cluster is saved as a representative Visual-Word (VW). Next, in the testing phase, the features extracted from the input image are compared with the VWs of each type and a visual-word histogram is created by voting for the nearest representative VW. Then CAD system classifies the testing image within a endoscopy movie (frame) by comparing the histogram made in the learning phase of each type with that of the testing image. Then it displays a supporting result for doctor as a "second opinion". In our software implementation, D-SIFT of Library VLFeat [3] is used for the feature extraction and Support Vector Machine (SVM) of LIBSVM [4] is used for type identification.

3. D-SIFT to VW Feature Transformation

The feature transformation is performed in both learning and testing phases (Fig. 2). First, at the learning phase, the local features of all the key points in all image areas (for example all 120x120 pixel area) are extracted from a training image dataset. Then D-SIFT features are transformed by the *hierarchical k-means clustering method* [5]. In our implementation, $k = 2$ and each clustering step divides the feature set to two sub-sets.

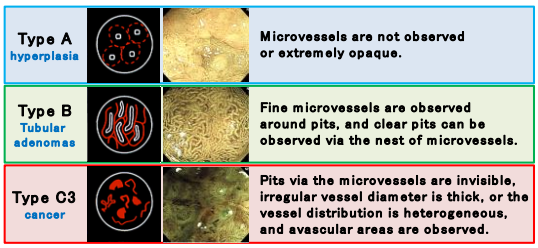


Figure 1. Narrow Band Imaging (NBI) magnification findings [1].

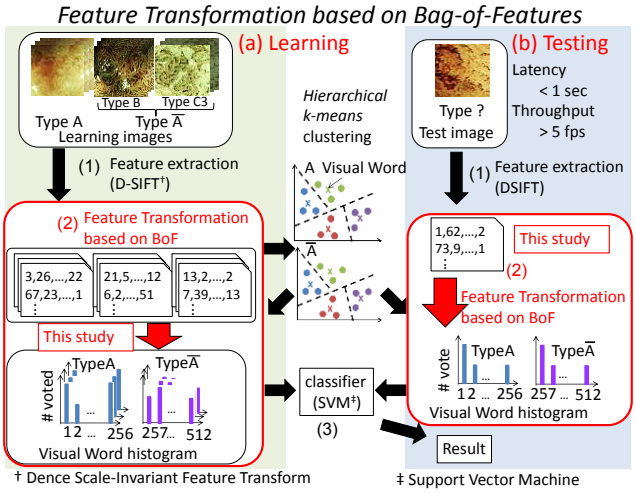


Figure 2. Computer-Aided Diagnosis System for Endoscopy Image.

The process repeats 8 times with the generated sub-set, making 256-VW and creating an *8-level binary searching tree*. Then, 256-dimension VW histograms are created by voting all the feature vectors at each key point extracted from the testing image to the nearest VW of each type among A, not A, B and C3. A 512-dimension VW histogram is then created by combining those two 256-dimension histogram into one. Figure 4 explains the outline of the feature transformation algorithm. A D-SIFT feature vector of each key point which is extracted as shown in Fig. 2 (a) (1) is compared with the representative feature vector in each node of the binary tree to find the nearest VW. This process repeats 8 times, until the input feature vector reaches to the leaf node, which is the sub-cluster (or VW in other word) that key point belong to. Hence, each input key point is voted on one of 256 VWs for each type. Since we have 3 types, all key points of the testing image will be voted on three 256-dimention VW histograms as shown in Fig. 2 (a) (2). Depend on the number of key points in the testing image, the VW histogram has different range of the maximum and the minimum. Hence, the histogram is normalized to the range [0, 1] by equation (1). The VW of i^{th} dimension is defined to VW_i, the normalized VW_i' is as follows.

$$VW_i' = \frac{VW_i}{\sqrt{\sum_{i=1}^{256} VW_i^2}} \qquad (1)$$

Then, the three 256-dimension histograms are combined together, making a 256-dimention x 2 types = 512-dimension VW histogram. This histogram is then used in Fig. 2 (a) (3) for testing phase. The same processing is done with the testing image by using the same VWs that are decided in learning phase. As a result, we get a 512-dimention VW histogram of the testing image. Also, there is the tradeoff of the distance in the distance comparison for feature vector classification. It helps to reduce both processing time and resources for VW histogram creation. In straight forward implementation, the computation of the denominator of equation (1) must wait until the VW histogram computation completes. We introduce a on-the-fly computation method for denominators of equation (1) to reduce the processing time in the hardware implementation. As a result, the waiting time for denominator computation in each dimension can be removed.

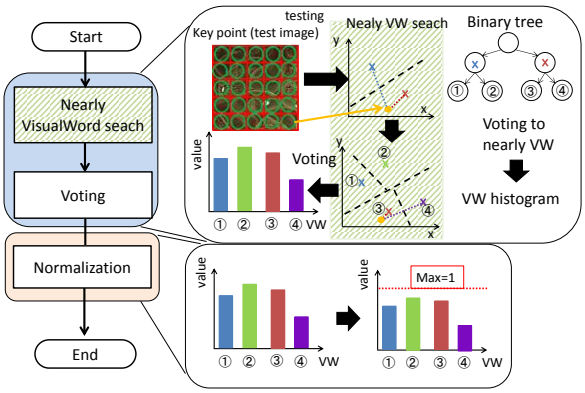


Figure 3. Feature transformation algorithm based on the Visual Word histogram.

4. The Proposed Feature Transformation Architecture

4. 1 Branch Processing Block

The branch processing block searches the nearest VW for the current inputted feature vector in the distance metric. The distance comparison procedure of the input feature vector and the 2 representative feature (left and right) vectors at each level of the binary search tree is shown in Fig. 3. At each level, result of the distance comparisons of left and right feature vectors are used to determine the next searching node on the binary tree (Fig. 5). If the distance difference $d_j \geq 0$, then the left child node is selected for the next searching, otherwise $d_i < 0$, the right child node is selected. The searching result at each level is represented by one bit for left or right. The combination of those results from 8 levels gives a 8-bit result, represents the VW number closed with the input feature vector among the 256 VWs given at the learning phase. Fig. 6 (a) shows the block design of the distance comparison block. Let f_j be the input feature vectors and l_j and r_j are defined as the representative feature vectors for left and right child nodes, respectively. Then the Euclidean distance is defined as $distance_{EU}(l_j, r_j) = (l_j, r_j)^2$ and the Manhattan distance is calculated $distance_{MH}(l_j, r_j) = |l_j - r_j|$, respectively.

4. 2 On-the-fly Normalization Computation

The Normalization block computes the L2 Norm ($\sum_{i=1}^{256} VW_i^2$) in the denominator of equation (1). The conventional method requires the normalization process waits until the calculation of this L2 Norm is finished. We improve the Normalization block to be able to calculate the L2 Norm during VW histogram creation by using partial sum.

Let M be the total number of feature vectors which will be voting in current input image. Let $N_j(VW_i)$ be the number of voting for the feature vector VW_i $(1 \leq i \leq 256)$ at the j^{th} voting. We define the partial sum of the L2 Norm for all VW_i at the j^{th} voting as equation (2).

$$S_j = \sum_{i=1}^{256} (VW_i)^2 \qquad (2)$$

When the $(j+1)^{th}$ voting is performed and the f_{j+1} is voted the Visual Word VW_k the partial sum of the L2 Norm for all VW_i at the $(j+1)^{th}$ voting can be calculated as following manner.

$$\begin{aligned} S_{j+1} &= \sum_{i=1}^{256} N_{j+1}(VW_i)^2 \\ &= \sum_{i=1, i\neq k}^{256} N_j(VW_i)^2 + (N_j(VW_k)+1)^2 \\ &= \sum_{i=1}^{256} N_j(VW_i)^2 + (2N_j(VW_k)+1) \\ &= S_j + 2N_j(VW_k) + 1 \end{aligned} \qquad (3)$$

From the equation (3), S_{j+1} can be calculated the partial sum S_j and the previous number of voting of $N_j(VW_i)$.

Because $N_0(VW_i)=0$ $(1 \leq i \leq 256)$, we can easily obtain $S_M = \sum_{i=1}^{256}(VW_i)^2$ by the above recurrent relation. So the partial sum can be calculated with the addition and shift operation on the fly of the voting. This is very suitable for hardware implementation without waiting time. The

improved architecture is shown in Fig. 6 (b). The previous voted number $N_j(VW_i)$ of the current Visual Word VW_i for f_j is read from the VW histogram memory and $N_j(VW_i)+1$ is stored to the memory. At the same time, $2N_j(VW_i) = N_j(VW_i) \ll 1$ is calculated by the left shift operation. In this way, we can obtain S_{j+1} by the previous partial sum on the fly. On the fly L2 Norm computation reduces 256 clocks for each SW in the image.

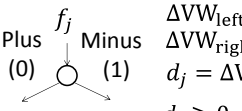
4. 3 Overview of the Architecture

The block diagram of the proposed high speed feature transformation architecture, which includes the *4-Parallel Binary Tree Search*, *4-Parallel Voting*, and *On-the-Fly Normalization*, and *SVM Scalling* blocks are shown in Fig. 7. First, the input feature vector is transformed to the VW histogram by searching for the nearest VW and voting to corresponding element in the 4-set of visual words which are used in SVM classification. Finally, the each VW histogram is normalized in the Normalization and Scalling blocks before sending to the SVM classifier.

5. FPGA Implementation and Evaluation

We have implemented feature transformation architechre with two distance metric (*Euclidian:EU* and *Manhattan: MH*) in the nearest neighbor search on FPGA, Altera Stratix IV (EP4SE530H35C2) device. The 4-parallel VW transformations are implemented for each sub-visual word, related with types A, B and C3. Their resource usages and processing time are shown in Table 1 for comparison. The DSP (Digital Signal Processing) block in Altera's FPGA is the dedicated block used to calculate the fixed-point multiplication in high speed. The type classifier with SVM [6], which received feature transformation result in Fig. 2, can optimize its processing speed by using multiple DSP blocks in parallel. Hence, reducing the number of DSP blocks used in VW feature transformation leaves more DSP resources for the critical SVM module. MH distance implementation saves 128 DSP blocks compared with EU distance. Figure 7 is our hardware (FPGA) and software test bench platform for D-SIFT-to-VW feature transformation. The platform receive the input image with capture board on PC via a HD-SDI cable. Then the D-SIFT feature extraction is processed on the PC, and send feature quantities to feature transformation module on the FPGA. Finally, the result of SVM module processed on the PC is displayed as the supporting image. Figure 8 explains the performance estimation about 60x60 scan window size feature transformation. Figure 9 shows the speed up estimation of the proposed feature transformation accelerator in comparison with software implementation for the SW size as large as 240x240 pixels, which has about 4,000 key-points. Without pipeline and parallel implementation, the hardware accelerator is 18 times faster than that of software implementation. 600 time faster can be achieved if 4-parallel 8-pipeline implementation is used. It guarantees the real time feature transformation (within 150 msec) for even the hard computation pyramid style hierarchical identification method, which contains nearly 2000 SW in 4 sizes in Fig. 10 [6]. The hardware accelerator removes the barrier of real-time processing in software implementation, which takes 64 sec to process the same amount of SWs for pyramid style hierarchical identification method. The proposed hardware can be used as a D-SIFT-to-VW feature transformation accelerator for binary tree searching engine beyond the endoscopic images. Also, we estimate the performance of the whole system with the estimation result of other modules, feature extraction [7], and type identifier module [6]. From the implementation results, throghtput is *16.7 fps* and latency is *60 msec*. So it is achivable about the real time and on-the-fly diagnostic support for clinical doctor (demand throghtput: >5 fps, latency: <1 sec).

$$distance\,(l_j, f_j) = (l_j - f_j)^2 \text{ or } |l_j - f_j|$$



$$\Delta VW_{left} = distance(l_j, f_j)$$
$$\Delta VW_{right} = distance(r_j, f_j)$$
$$d_j = \Delta VW_{left} - \Delta VW_{right}$$
$$d_j \geq 0 \;\; right \qquad d_j < 0 \;\; left$$

f_j : j^{th} input feature vector
l_j : j^{th} feature vector of left cluster
r_j : j^{th} feature vector of right cluster

Figure 4. Calculation at each level of binary tree.

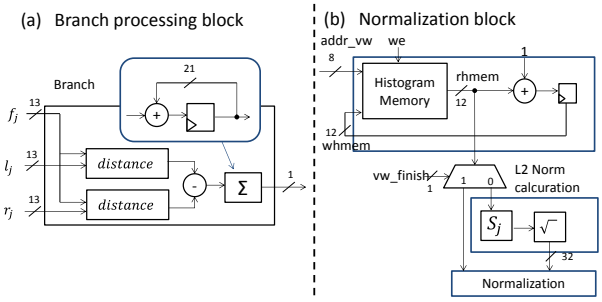


Figure 5. Processing blocks.
(a) Branch processing block, (b) Normalization block.

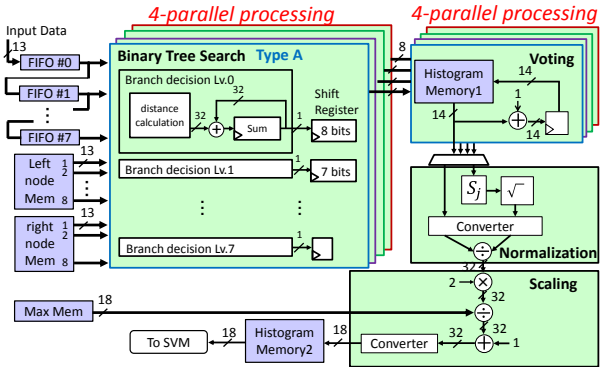


Figure 6. The proposed feature transformation architecture.

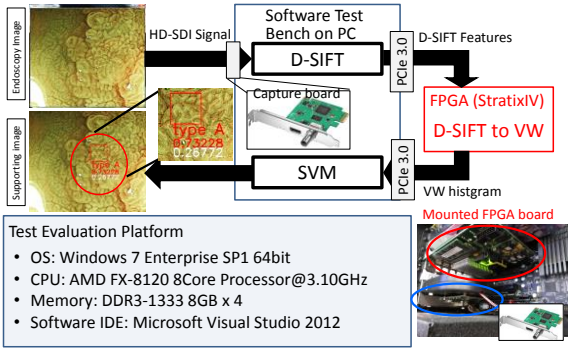


Figure 7. Evaluation platform with hardware and software co-design system.

6. Conclusion

In this paper, we have proposed the FPGA based hardware accelerator for D-SIFT to VW feature transformation in real-time CAD system. From the implementation result on FPGA evaluation board, it is very promising to use at an actual medical clinic in the view points of supporting quality and real-time processing performance. The processing time for 240x240 scan window can be as fast as 0.15 msec@100 MHz (with 4 parallel and 8 pipeline) and it is about 600 times faster than that of software implementation (90 msec). It guarantees that feature of 2000 SW in 4 SW sizes in complexed pyramid style hierarchical identification method can be processed in 150 msec. Future work includes the development of the whole CAD system including our D-SIFT architecture [7] and our SVM architecture [6] in one FPGA board.

Acknowledgment

Part of this work was supported by Grant-in-Aid for Scientific Research (B) JSPS KAKENHI and JSPS Fellows, Grant Numbers 26280015 and 16J06130, respectively, and was with the help of a grant by Chugoku Industrial Innovation Center. The FPGA design tools in this work have been supported by the Altera University Program and the Mentor Graphics Higher Education Program.

References

[1] H. Kanao, et al., "Narrow–band imaging magni fication predicts the histology and invasion depth of colorectal tumors," Journal of Gastrointestinal Endoscopy, vol. 69, no.3, pp. 631-636, 2009.

[2] T. Tamaki, et al., "Computer-aided colorectal tumor classification in NBI endoscopy using local features", Medical Image Analysis, Vol. 17, No. 1, pp. 78-100, 2013.

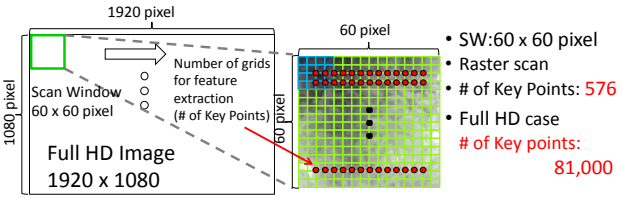
[3] A. Vedaldi, and B. Fulkerson, "Vlfeat: an open and portable library of computer vision algorithms," http://www.vlfeat.org/

[4] Chin-Chung Chang, Chin-Jen Lin, "Livsvm – a library for support vector machins," http://www. csie.ntu.edu.tw/~cjlin/ libsvm/

[5] Nister, and H. Stewenius, "Scalable recongnition with a vocabulary tree", Proc. of the IEEE Computer Vision and Pattern Recognition (CVPR2006), pp. 775-781, 2006.

[6] T. Okamoto, et al., "Image segmentation of pyramid style identifier based on support vector machine for colorectal endoscopic images", Proc. of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp.2997 – 3000, Italy, August 2015.

[7] T. Koide, et al., "A hardware accelerator for bag-of-features based visual word transformation in computer aided diagnosis for colorectal endoscopic images", Proc. of the 31th International Technical Conference on Circuits/Systems, Computers and Communications 2016, July 2016 (to appear).



FPGA Board	PROCe IV 530-A (GiDEL)
Memory	On-borad DDRII 512MB DDRIISODIMM 1GB x 2
Installed FPGA	Altera Stratix IV EP4SE530H35C2
# of FPGAs	1
Host Interface	PCI-Express Gen 3.0

Available Resources

Altera Stratix IV Available	
# ALUTs	424,960
# registers	424,960
# Total RAM bits	21,233,664
# 18 x 18 DSPs	1,024

Figure 8. Performance evaluation for Full-HD endoscopy image with 60x60 scan window size.

Table 1. FPGA implementation results.

Resources	Usage/Available (Utilization)	
	Euclid distance	Manhattan distance
Number of ALUTs	47,812 / 424,960 (11%)	47,680 / 424,960 (12%)
Number of Register	16,368 / 424,960 (4%)	14,220 / 424,960 (3%)
Total RAM [bit]	962,560 / 21,233,664 (8%)	962,560 / 21,233,664 (8%)
Multiplier (# of DSP blocks)	164 / 1,024 (16%)	36 / 1,024 (4%)

Performance	Demand	Euclid distance	Manhattan distance
Max Operating Frequency		116.21 MHz	105.85 MHz
Latency @100MHz	< 1 sec	60 msec@100MHz	60 msec@100MHz
Throughput @100MHz	> 5 fps	16.7 fps	16.7 fps

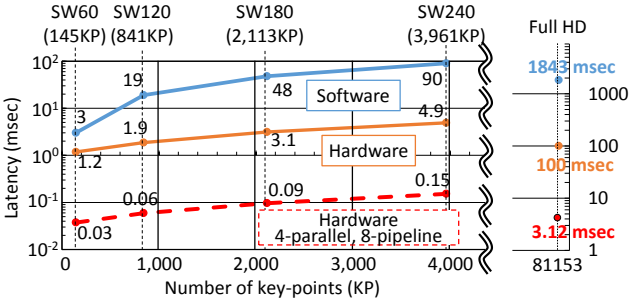


Figure 9. Performance estimation for Full-HD endoscopy image with 4-parallel 8-pipeline implementation.

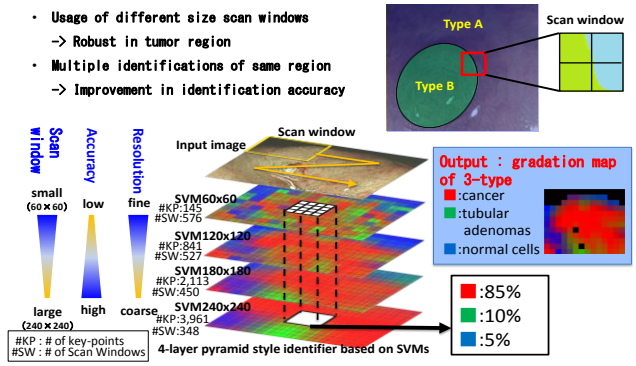


Figure 10. Concept of the image segmentation method based on bottom-up hierarchical (pyramid style) SVM identifiers [6].