Extended Abstracts of the 2015 International Conference on Solid State Devices and Materials, Sapporo, 2015, pp144-145

Associative Memory for Nearest Neighbor Search with High Flexibility of Reference-Vector Number Due to Configurable Dual-Storage Space

Fengwei An, Keisuke Mihara, Shogo Yamasaki, Lei Chen, and Hans Jürgen Mattausch

Hiroshima University, 1-3-1 Higashi-Hiroshima, Hiroshima 739-8530, Japan
Phone: +81-82-424-5730; E-mail: anfengwei@hiroshima-u.ac.jp

Abstract

In this paper, a digital word-parallel associative memory architecture with reconfigurable dual-storage space and flexible number of reference vectors is reported for nearest squared Euclidean distance search, applying a clock-counting concept. The clock-based minimal-distance searching is implemented by weighted frequency dividers and achieves high classification speed, good area-efficiency and low power dissipation. Switching circuits, located between vector components, enable scalability of the reference feature-vector dimension. To avoid the limitation of reference-vector number, a pipeline storage with dual SRAM cells for each unit and an intermediate winner control circuit are designed to extend the applicability. A test chip in 180 nm CMOS technology, which has 32 rows, 4 elements in each row and 2-parallel 8-bit dual-components in each element, achieves low power dissipation of 61.4 mW (at 45.58 MHz clock frequency and 1.8 V supply voltage).

1. Introduction

The main limitation of the usage of a nearest-neighbor-search (NNS) classifier [1], which realizes one of the most basic algorithms in pattern recognition to classify unknown samples, is the high computational costs of the minimal distance searching in traditional software implementations. Pattern recognition in mobile or wearable devices has attracted much attention for a wide range of applications. Thus, special-purpose hardware implementations for NNS [2-4] have been proposed which significantly outperform software implementations with respect to recognition speed. In general, a hardware implementation achieves also higher energy efficiency than a comparable sequential software implementation.

The main contributions of this paper can be attributed to 3 aspects: (a) Mapping of the minimal squared Euclidean distance search into a clock-based time domain for high speed. (b) Flexibility in dimensionality and number of reference vectors due to programming switches for match-signal connections between clocked search circuits. (c) Further increased flexibility in the number of reference feature vectors by dual SRAM storage with high writing bandwidth and control circuitry with intermediate winner-data storage to enable pipelined search among a freely selectable number of reference-vector blocks.

2. Clock-based nearest-neighbor search (NNS)

The NNS classifier assigns an unknown input to the class of the most similar reference vector in terms of the squared Euclidean distance (SED) of eq. (1). Traditionally,

$$SED = \sum_{j\le d}^{0<i\le n} \left(REF_{ij} - IN_j \right)^2 \qquad (1)$$

the brute-force smallest SED search among n reference vectors requires O(dn) time for an unknown d-dimensional input feature vector. Rather than adders and comparators, weighted value counters (WVCs) are applied for minimal-distance searching with lower power dissipation and smaller chip area. Each bit of a WVC consists of a frequency divider (with 21 transistors [5]), a multiplexer, a XOR gate, an AND gate, and a transmission gate. The distance evaluation unit (DEU) uses the WVC and can achieve an only linear worst-case search-clock-number increase according to 2N(d+1)-1 for the worst-case nearest SED search [3].

3. Programming switches for dimension flexibility

A reconfigurable associative memory (RASM) concept is implemented as an alternative of the dimension extension circuit [4] to improve the flexibility in both reference-vector dimension and number. The RASM consists of elements arranged in R rows and M columns. Each element contains p-word SRAM cells, p-vector distance computing units (DCUs) and one DEU. The example of Fig. 1 with R=4, M=3 and 2-dimensional elements can be configured into 6 associative-memory variations (1-ref., 24-dim.; 2-ref., 12-dim.; 3-ref., 8-dim.; 4-ref., 6-dim.; 6-ref., 4-dim.; 12-ref., 2-dim.) for reference-vector number and dimension by placing switches between these elements. In the general case of R rows, M columns and d-dimensional feature vectors, (R×M×p)/d vectors can be processed in parallel to find

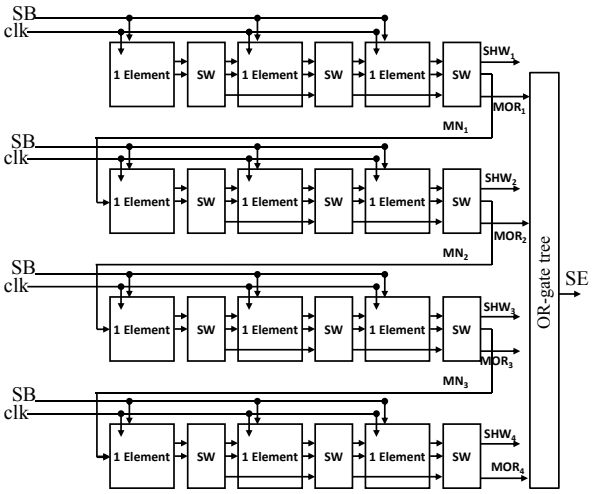


Fig.1. Overview diagram of the reconfigurable associative memory architecture for NNS with a 4-row, 3-column example for the arrangement of basic elements (≥1 vector component).

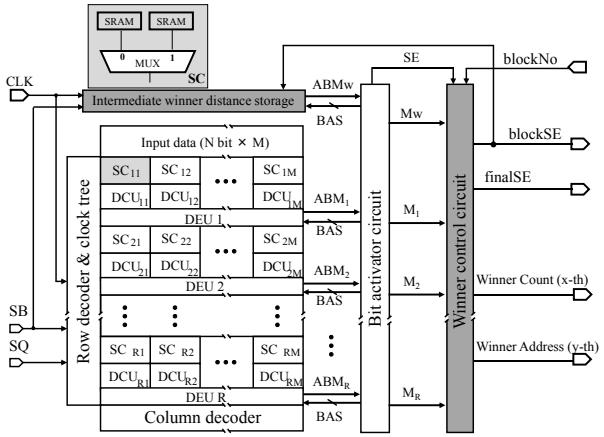


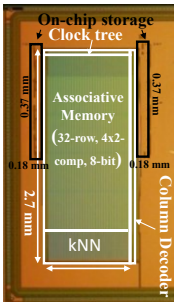
Fig.2. Associative memory with dual-SRAM storage, intermediate winner control circuitry, as well as parallel search and write capability for flexibility of the reference number.

the minimal distance by appropriately reconfiguring the switches between the elements. SRAM cells are associated with the functional logic circuits for high efficiency. The configuring signal of the multiplexing switch should be initialized by pre-stored information in the memory. The switches, which mainly control the connection of the match signals for every element, OR tree, and final winner-signal reading, provide thus high flexibility for the number of reference vectors and their dimensionality.

4. Pipelined exchange of stored reference vectors

To avoid the limitations of providing large on-chip associative memory, a pipelined storage exchange and additional control circuitry for intermediate winner data are designed as shown in Fig. 2. The dual-SRAM storage in each element, for parallel partial minimum-distance search and writing of new reference data, further increases the flexibility with respect to the reference-vector number. This means it also increases the capability of satisfying the needs of multiple applications.

At the beginning of the search process, the intermediate winner distance is initialized to the maximum possible distance and the first block of reference vectors is pre-stored in one the dual SRAM storages (SCs). Then, the minimum distance is searched in parallel within the first reference-vector block by the circuitry for distance mapping on clock number. Simultaneously, the second reference-vector block is stored in the other part of the dual SRAM storage. To obtain sufficient writing bandwidth for approximately equal time delay of second block writing and first-block search, multiple rows (2 in the test chip) are written in parallel. After the distance searching is completed (SE is asserted), the local minimal distance is transferred to the intermediate winner-distance storage circuit and the block-number (blockNo) counter of the winner-search control circuit is increased by 1. The winner search continues sequentially for the reference-vector blocks and the intermediate winner-distance storage circuit is updated each time when a new local minimum is found. After the last reference-vector block is searched, the finalSE signal is asserted and block number (X), row number (Y) and distance of the winner are retrieved from winner-search control circuit.



Distance metric	Euclidean
Technology	180 nm
Supply voltage	1.8 V
Parallelism	256
Total Power (mW)	61.4 (45.58 MHz)
Power for search (mW)	11.9
Search time (average case)	4.38 µs
Power-delay product (nJ)	52.12
Area (mm^2)	3.88

Fig.3. Photomicrograph of the fabricated test-chip which includes reconfigurable word-parallel associative memory, dual-storage SRAM and highly-flexible kNN-classification.

5. Experimental results and conclusion

The fabricated test chip (Fig. 3) for the proposed RASM architecture in 180 nm CMOS technology has 32 rows, 4 elements in each row and 2 vector components of 8-bit per element. In particular, each element in Fig.1 has one DEC and one DEU with 24 bits that can extend the search capability to 2048-dimensional feature vectors. 2 reference vectors can be written in parallel to enable high-bandwidth exchange of stored reference-vector blocks. The measured delay within the DEU until the output of the WVC for the currently evaluated distance bit is about 0.59 ns. The path delay through the AND gates in the match-signal path of all 4 DEUs, the programming switches between DEUs, and the OR-gate tree is measured at 7.79 ns. Consequently, the maximum working frequency is determined to be about 120 MHz. The power dissipation is measured to be 61.4 mW (at 45.58 MHz, 1.8 V supply voltage) using the on-chip ring oscillator and average search configurations for the distance computing units DEC and DEU. The distance computation by DECs and DEUs, which requires only 8 clock cycles in the case of 8-bit words, uses more than 80% of the total power dissipation. Consequently, the minimal SED search consume only about 11.9 mW. In comparison to our previous work in [4], the match signal in DEU has 0.2 ns increased delay per bit due to the reconfigurable switches. However, the work reported here has achieved much higher flexibility and applicability.

In conclusion, besides the designed flexibility for dimension and number of stored reference vectors, high additional efficiency in speed performance, area consumption and power dissipation could be demonstrated through the experimental results.

Acknowledgements

This research was supported by grant 25420332 from the ministry of Science and Education, Japan. The VLSI-chip was fabricated through the chip fabrication program of VDEC, the University of Tokyo in collaboration with, Rohm, Synopsys, and Cadence.

References

[1] T.M. Cover, et al., *IEEE Trans. on Info. Theory*, 13 (1967) 21-27.
[2] T. Akazawa, et al., *Jpn. J. Appl. Phys.*, 53 (2014) 04EE16.
[3] S. Sasaki, et al., *ESSCIRC*, (2012) 185-188.
[4] F. An, et al, *IEEE CICC*, (2014), DOI: 10.1109/CICC.2014.6946096.
[5] E. A. Vittoz, et al., *IEEE JSSC*, 7 (2), (1972) 100-104.