An Energy-efficient Non-volatile In-Memory Accelerator for Sparse-representation based Face Recognition

Yuhao Wang*, Hantao Huang*, Leibin Ni*, Hao Yu*, Mei Yan*, Chuliang Weng†, Wei Yang†, Junfeng Zhao†
*School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
†Shannon Laboratory, Huawei Technologies Co., Ltd, China

***Abstract*—Data analytics such as face recognition involves large volume of image data, and hence leads to grand challenge on mobile platform design with strict power requirement. Emerging non-volatile STT-MRAM has the minimum leakage power and comparable speed to SRAM, and hence is considered as a promising candidate for data-oriented mobile computing. However, there exists significantly higher write-energy for STT-MRAM when compared to the SRAM. Based on the use of STT-MRAM, this paper introduces an energy-efficient non-volatile in-memory accelerator for a sparse-representation based face recognition algorithm. We find that by projecting high-dimension image data to much lower dimension, the current scaling for STT-MRAM write operation can be applied aggressively, which leads to significant power reduction yet maintains quality-of-service for face recognition. Specifically, compared to a baseline with SRAM, leakage power and dynamic power are reduced by 91.4% and 79% respectively with only slight compromise on recognition rate.**

I. Introduction

Applications such as object and face recognitions involve classification or convex optimization with high complexity between test image and numerous images in database [1], [2], which have posed strict power requirement for both memory and logic components on mobile platforms where battery lifetime is limited.

To maintain and access image data in memory, substantial leakage power will be experienced for the conventional volatile SRAM. In [3], aggressive voltage scaling is applied to volatile memory that can reduce the leakage power. Emerging non-volatile memory technologies such as spin-toque-transfer magnetic random access memory (STT-MRAM) provides low leakage option for the future data-oriented computing. However, there exists significantly higher write-power for STT-MRAM compared to SRAM. For example, the work in [4] reports that the 5nJ write-energy for STT-MRAM is more than 6x higher than the 0.8nJ write-energy for SRAM.

Moreover, the logic accelerator for face recognition algorithm involves the image data classification or optimization in significantly high dimension when operating raw image data directly, thus is extremely power and bandwidth demanding. The conventional approach is to extract features of image, reduce feature dimensions and then perform the recognition based on dimensionally reduced features. The feature extraction will still be computation intensive with loss of accuracy. For example, the biometrics based feature extraction searches hundreds of features with various sizes of box determined iteratively throughout the images. Moreover, the deployed dimension reduction by PCA will introduce additional computational cost of energy. Recently, the compressive sensing based sparse-representations have empowered the possibility of energy efficient data analytics such as face recognition [5]. Assume the test sample is a linear combination of a few images of same person from the database, thus the correlation matrix can be sparse, which can be solved by $\ell 1$-norm minimization.

In this paper, we have introduced an energy-efficient non-volatile in-memory accelerator for sparse-representation based face recognition algorithm. Based on the STT-MRAM with the minimum leakage power, aggressive current scaling is applied for STT-MRAM to reduce the write operation energy with slight errors incurred. Moreover, the feature extraction and dimension reduction are performed simultaneously by random projection, which has much lower complexity than the conventional feature extraction and dimension reduction methods. We find that when the feature projection is performed properly, the errors caused by current scaling can be well tolerated. An implicit OMP classifier is deployed for sparse-representated images recognition without the need of energy expensive recovery [6]. The in-memory architecture [7], [8] is adopted to perform data compression and service together with sparse data storage. Experimental results show that significant power reduction can be achieved with only slight compromise on the face recognition rate.

The rest of this paper is organized as follows. Section II introduces the proposed energy-efficient face recognition hardware architecture with module details in Section III. Experiment results are presented in Section IV with conclusion in Section V.

II. System Overview

There exists some challenges for the implementation of sparse-representation based algorithm from hardware perspective. At front end, there exists significant data exchange between external processor and memory when perform feature extraction and dimension reduction, and the memory bandwidth may limit the overall performance. In addition, to maintain and frequently access data can lead to high power consumption on memory. At back end, to represent the image

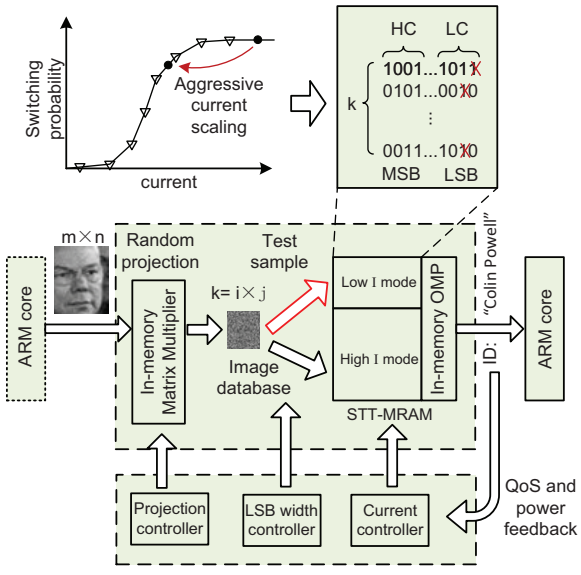


Fig. 1. The system overview of the proposed energy efficient face recognition architecture

database by sensing matrix and find sparsest representation is also extremely memory and computation demanding.

The energy efficient non-volatile in-memory accelerator for the sparse representation base face recognition application is proposed and illustrated in Fig. 1. The proposed architecture features the following designs to overcome aforementioned challenges,

- The non-volatile STT-MRAM is used to keep leakage power minimal.
- The STT-MRAM is equipped with in-memory logic as interface between memory and external processor.
- An in-memory matrix multiplier is designed to perform random projection, which is used to reduce the dimensionality and extract features of both the database image and test sample. The recognition is then performed in low dimension based on features instead of pixels.
- The STT-MRAM is operated in low current mode by aggressive current scaling to reduce dynamic power. The aggressive current scaling shifts the safe working point of STT-MRAM to a point that has a lower writing current but with slight probability of incurring errors. The errors can be tolerated by the feature based recognition.
- A in-memory OMP logic is deployed to respond to sparse-representation based recognition request from external processor.
- QoS orientated parameter tuning for energy efficiency optimization that can be implemented in offline fashion or online hardware module.

III. Non-volatile In-memory Accelerator

A. Random Projection for STT-MRAM Input Reduction

As the image data is commonly of high dimension, 640×480 for example, it is necessary to reduce the dimensionality of the original data to before processing. Among various methods like down-sampling or PCA, random projection is a computationally efficient yet effective means of dimension reduction.

The random projection casts points from high-dimensional into low-dimensional Euclidean space, and the distances between the points are approximately preserved. Such transformation can be denoted by the projection matrix R, and it satisfies the conditions that all elements follow Bernoulli or Gaussian distribution and all columns have unit lengths. The image data A from d-dimension ($d = m \times n$) is projected to k-dimension A' by multiplying a random $k \times d$ matrix R,

$$A'_{k\times 1} = R_{k\times d} A_{d\times 1} \tag{1}$$

As STT-MRAM is power hungry for write operation, random projection can reduce the total number of write operations, thus reduce the memory dynamic power. The solver logic workload can also be greatly alleviated, as the solver to Eq. 5 has a approximate time complexity of $\sim O(n^3)$.

Besides the dimension reduction, random projection extracts the image features as well. The conventional feature extraction uses holistic Eigenfaces [1], or biometrics patches [2] as features, but such approaches involves additional computations compared to the random projection. The random projected vectors are proven to be valid features for face recognition as suggested in [5].

B. Current Scaling for Low Energy STT-MRAM Operation

For conventional volatile memory, voltage scaling can cause three types of errors: write error, read error, and hold error. For STT-MRAM, there is no hold error as it does not require to be powered for holding data; and the read current is usually far smaller than write current, so the current scaling will not apply for read operation. The write operation of non-volatile memories, on the other hand, requires large current that produces enough energy to reverse the magnetization in target cell.

In fact, the switching probability of STT-MRAM cell depends on both applied current amplitude as well as pulse duration. As indicated in [9], with a fixed pulse width, the switching probability of STT-MRAM cell under different writing current follows,

$$p(I) = 1 - exp(-\frac{I-\alpha}{I_0}) \tag{2}$$

where I_0 is the critical current and α the fitting parameter.

When a bit fails to switch, the written data will deviate from desired value. Assume the target binary data to write is Y, and the actual data written is $\hat{Y}$ when the error occurs in the n_{th} bit, the relation can be formulated as

$$\hat{Y} = Y + \begin{cases} e^+, & Y_n = 0 \\ e^-, & Y_n = 1 \end{cases} \tag{3}$$

When the n_{th} is 0 then the error can only occur when 1 is mis-written, which introduces positive error e^+; and when the n_{th} is 1 then a negative e^- will be incurred if writing 1 is unsuccessful. The positive and negative errors subject to the following probability distribution,

$$e^{\pm} \sim P(e^{\pm}|I) = \begin{cases} 1-p(I), & e^{\pm} = \pm 2^n \\ p(I), & e^{\pm} = 0 \end{cases} \tag{4}$$

While the errors in least significant bits (LSB) may bring small deviation of original data, those in most significant bits (MSB) will bring significant impact which is considered unacceptable. Therefore, the current scaling will only be applied to LSB while the data integrity of MSB will not be compromised.

C. OMP for STT-MRAM Output Reduction

The sparse representation based face recognition, discussed in the work [5], can be interpreted from a compressive sensing perspective. Assume the test sample is a linear combination of a few images of same person from the database, thus the correlation matrix can be sparse, while the training samples in database can be viewed as sensing matrix. In other word, the test sample has a sparse representation on the domain of training samples. As such, the face recognition can be performed by solving the following equation,

$$\begin{aligned} \arg\min_{x\in\mathbb{R}^N} \quad & \|x\|_1 \\ \text{subject to} \quad & y = \Phi x \end{aligned} \tag{5}$$

where y is the feature after random projection, Φ is the image database and x is the solution sparse coefficient matrix.

The commonly used solvers for convex problem for Eq. 5 are of high complexity. They are either resource consuming for the general purpose processor or hard to implement as hardware. Beside the convex problem solvers, there are also greedy algorithms such as matching pursuit (MP) and orthogonal matching pursuit (OMP) [10].

In this work, the OMP solver is adopted to find the sparsest solution for Eq. 5 with hardware implementation. The key idea is that Y is composed from the columns of sensing matrix Φ and therefore, we can calculate the contribution of each column of the sensing matrix Φ by projecting Y to sensing matrix Φ. The largest correlation X_i between Y and columns Φ_i is selected and the residual is calculated by $Y_{res} = Y - X_i\Phi_i$. This procedure will repeats until the residual is insignificant. This method can be implemented in the hardware efficiently [11] and by using square-root-free Cholesky factorization,

the complexity can be reduced from $O(MK^3)$ to $O(MK^2)$, where K is the sparsity.

D. QoS orientated Energy Efficiency Optimization

As the dynamic write energy is the main concern for STT-MRAM, the goal is to optimize the dynamic write energy while maintain the quality of service (QoS), defined as face recognition success rate. The three parameters to consider are the scaled current level I, LSB width w, and the dimension reduction ratio γ. The parameter $\gamma = \frac{k}{d}$ is defined as dimension reduction ratio between the original image data in d-dimension and the projected feature vector in k-dimension, and w depicts the width of LSB which are operated in low current mode. The I determines the energy consumed per write operation, w and γ decide the number of operations necessary. Therefore, the total write energy consumption with all three parameters for the face recognition application can be defined as,

$$E_w \propto \gamma((N-w)\cdot {I_s}^2 + w\cdot I^2) \tag{6}$$

where I_s is the normal case write current, and N is the total width of feature data.

Consider that QoS is obtained by statistically testing dozens of images by solving Eq. 5, to find optimized power in the constructed three dimensional design space is extremely time consuming. In this paper, the design space exploration is done in two steps. Firstly, we have examined the relations among all three parameters and their impact on QoS with large dynamic ranges. The core sub-space with largest descent is then identified, in which the optimal solution must exist. Secondly, within the confined sub-space a greedy coordinate searching method is deployed to find the close-optimal solution with small dynamic range.

The ***greedy coordinate*** searching method is applied to locally optimize the three dimensional sub-space. As there are three parameters in the design space, a three dimensional coordinate can be constructed with I, γ and w as axes. The searching method starts from empirical given point, and search in an iterative way. Within each round, the point has three directions to move, and the one yields highest power saving while meets the QoS requirement will be favored as move decision. The iterations continue until none direction can provide a further reduction of power under the given QoS constraint.

IV. Simulation Results

A. Simulation Setup

As the proposed architecture aims for power efficient yet accurate face recognition, the face recognition success rate is adopted as the measurement of the system service performance, namely QoS. The QoS is calculated based on randomly selected 100 images from the LFW face database [12]. All images are down-sampled to 40×50 before the experiment.

The evaluation of the proposed architecture is done in three steps. Firstly, aggressive current scaling and random projection are evaluated separately to examine the trade-offs between QoS and power. Secondly, the two techniques are applied jointly, and the design space is explored to determine the best combination of the design parameters. Thirdly, based on the optimized design parameters, hardware performance is evaluated both from memory and logic perspective. For the memory performance assessment, CACTI [13] is used to obtain the SRAM performance in the baseline system and NVSim [14] is employed to get the performance of STT-MRAM. The technology node of 65nm is adopted for both memory and logic designs.

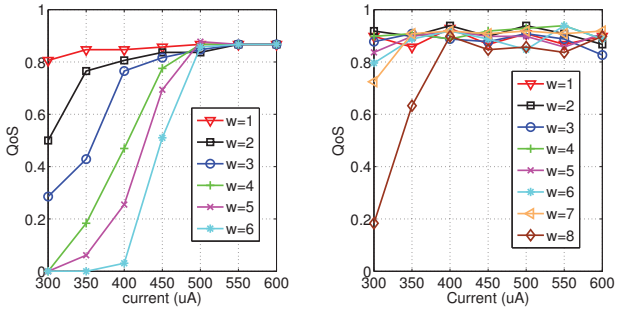


Fig. 2. The face recognition success rate as QoS under different scaled STT-MRAM writing currents (a) without dimension reduction ($\gamma = 100\%$) and (b) with dimension reduction ratio $\gamma = 80\%$. In both figures, the w denotes the width of LSB that aggressive current scaling is applied to.

B. Design Space Exploration

1) Aggressive Current Scaling: In this part, we will investigate the relations between QoS and aggressive current scaling parameters, i.e. the least significant bits (LSB) width w and writing current I. To characterize the switching errors of STT-MRAM, the model parameters are set with I_o = 100 and α = 200, which is fitted with measurement data in [9] . The switching probabilities at various writing current levels is shown in Table I.

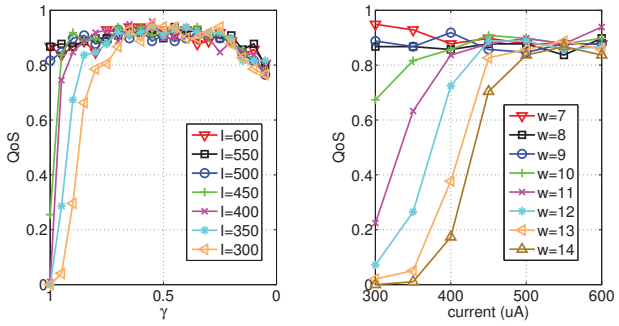
Under random projection without dimension reduction ($\gamma = 1$), the QoS at different scaled current levels with varying LSB width w are presented in Fig. 2(a). As indicated in Table I, the error probability increases as the current is scaled down. Therefore, the QoS will decrease as more errors are incurred for low current cases. This is not obvious for a small LSB width w, but significant for large w. Although more energy saving can be anticipated, a large number of w means not only more bits may suffer from errors, but also the magnitude of the errors will increase exponentially. For example, with $w = 1$ the current can be half scaled while the QoS can be almost preserved even approximate one third of the projected features have suffered errors of ± 1 on the last bit. For $w = 6$ at $I \leq 350$, more than one tenth of the projected features will incur errors ranging from minimal ± 1 to maximal $\pm 2^6 - 1$. The deviation is substantial to damage the features, and result shows no successful identification is achieved.

For random projection with $\gamma = 0.8$, the QoS against current results are illustrated in Fig. 2(b). The system with $\gamma = 0.8$ shows better performance compared to their counterparts in Fig. 2(a). For example, in system with $w = 6, \gamma = 0.8$ the current can be half scaled and still provide ideal QoS, yet in system with $w = 6, \gamma = 1$ the current can only be scaled from 600 to 500. The system performance under $w = 8, \gamma = 0.8$ is comparable to that of $w = 2$ without dimension reduction.

2) Random Projection: Figure 3(a) demonstrates the QoS at different dimension reduction ratio. It can be observed that the QoS peaked while γ is around 0.5. For high γ end, the QoS is susceptible to

TABLE I
THE STT-MRAM CELL SWITCHING PROBABILITY UNDER DIFFERENT I

current (μA)	550	500	450	400
P_{switch}	1.0000	0.9999	0.9981	0.9817
current (μA)	350	300	250	200
P_{switch}	0.8946	0.6321	0.2212	0.0000



(a) (b)

Fig. 3. (a) Random Projection with different dimension reduction rate γ at $w = 7$ (b) Low γ end ($\gamma = 0.2$) tolerance of different w.

errors, as the features are poorly preserved when errors are incurred. At the low γ end, the QoS damps in a slower slope than that in the high γ end. It can be inferred that the degradation is caused by over-reduction of the dimension instead of errors, as the QoS shows very low dependency on I and errors. In other words, random projection with low γ is more robust to errors. To verify this, Fig. 3(b) shows the QoS under γ with increased w, and it can be observe that while random projection high γ can only endure $w \leq 7$, low γ end is robust to errors with $w \leq 11$.

C. Accelerator Power Performance Comparisons

In this part, we compare the proposed accelerator architecture with the baseline system. The baseline system is a typical SRAM for storage and general purpose processor for computation. To fit the LFW face database size, the memory size is configured to be $16MB$. The leakage power P_l, normalized writing energy $\tilde{E}_w$, and design area are selected as metrics for memory design.

For the proposed memory design, the parameters $w = 7$, $I = 450$, and $\gamma = 0.1$ are used, and the memory is configured with four different schemes: 1). non-volatile STT-MRAM only, 2). STT-MRAM with aggressive current scaling, 3). STT-MRAM with random projection, and 4). STT-MRAM with both aggressive current scaling and random projection. The memory performance comparison is listed in Table II. Compared to the SRAM baseline memory, all four schemes with STT-MRAM deployed have reduced the leakage power P_l by 91.4% due to the non-volatility of STT-MRAM. For the dynamic writing operation energy, the STT-MRAM has 3.8x higher writing energy compared to baseline design. This is due to the high current and energy requirement to reverse the magnetization in each cell. With the help of aggressive current scaling, $\tilde{E}_w$ can be improved by 44% for NVM, yet still 2.13x higher than baseline design. Consider the STT-MRAM with random projection deployed, $\tilde{E}_w$ is reduced by 59% compared to the baseline design. And when both techniques are applied simultaneously, the $\tilde{E}_w$ is only one-fifth of the baseline. For memory area, all four schemes are 2.5x times smaller than SRAM. This is because of the cell efficiency of 1T1MTJ in STT-MRAM compared to the 6T structure of SRAM.

TABLE II
THE PERFORMANCE COMPARISON FOR MEMORY

Scheme	P_l	$\tilde{E}_w$	Area
baseline	7.17mW	1.00	$86.31mm^2$
NVM	614.55μW	3.80	$33.83mm^2$
NVM+ACS†		2.13	
NVM+RP‡		0.41	
NVM+ACS+RP		0.21	

† aggressive current scaling
‡ random projection

V. Conclusion

With the use of STT-MRAM, an energy-efficient non-volatile in-memory accelerator is introduced for sparse-representation based face recognition. In order to achieve the low power consumption yet to process the large image data in face recognition, the developed non-volatile in-memory accelerator has the following features. Firstly, random-projection based sampling is deployed to compress large-volume image data from high-dimension to low-dimension. Moreover, STT-MRAM current more is aggressively scaled down under the quality-of-service constraint. Experimental results show that for the memory part, the leakage power and dynamic power are reduced by 91.4% and 79% respectively with only slight compromise on the face recognition rate, when compared to a the SRAM based design.

References

[1] J. Zhang and et al., "Face recognition: eigenface, elastic matching, and neural nets," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1423–1435, 1997.
[2] P. J. Phillips and et al., "The feret evaluation methodology for face-recognition algorithms," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 10, pp. 1090–1104, 2000.
[3] D. Bortolotti and et al., "Approximate compressed sensing: ultra-low power biosignal processing via aggressive voltage scaling on a hybrid memory multi-core processor," in *Proc. of ISLPED*. IEEE/ACM, 2014, pp. 40–45.
[4] G. Sun and et al., "A novel architecture of the 3d stacked mram l2 cache for cmps," in *High Performance Computer Architecture, 2009. HPCA 2009. IEEE 15th International Symposium on*. IEEE, 2009, pp. 239–249.
[5] J. Wright and et al., "Robust face recognition via sparse representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 2, pp. 210–227, 2009.
[6] L. Zhao and et al., "An improved auto-calibration algorithm based on sparse bayesian learning framework," *Signal Processing Letters, IEEE*, vol. 20, no. 9, pp. 889–892, 2013.
[7] Y. Wang and H. Yu, "An ultralow-power memory-based big-data computing platform by nonvolatile domain-wall nanowire devices," in *Low Power Electronics and Design (ISLPED), 2013 IEEE International Symposium on*. IEEE, 2013, pp. 329–334.
[8] Y. Wang, H. Yu, D. Sylvester, and P. Kong, "Energy efficient in-memory aes encryption based on nonvolatile domain-wall nanowire," in *Design, Automation and Test in Europe Conference and Exhibition (DATE), 2014*. IEEE, 2014, pp. 1–4.
[9] L.-B. Faber and et al., "Dynamic compact model of spin-transfer torque based magnetic tunnel junction (mtj)," in *Design & Technology of Integrated Systems in Nanoscal Era, 2009. DTIS'09. 4th International Conference on*. IEEE, 2009, pp. 130–135.
[10] Y. C. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Signals, Systems and Computers, 1993.* IEEE, 1993, pp. 40–44.
[11] F. Ren and et al., "A square-root-free matrix decomposition method for energy-efficient least squares computation on embedded systems," *IEEE Embedded Systems Letters*, vol. Early access, 2014.
[12] G. B. Huang and et al., "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.
[13] S. J. Wilton and N. P. Jouppi, "Cacti: An enhanced cache access and cycle time model," *Solid-State Circuits, IEEE Journal of*, vol. 31, no. 5, pp. 677–688, 1996.
[14] X. Dong and et al., "Nvsim: A circuit-level performance, energy, and area model for emerging nonvolatile memory," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 31, no. 7, pp. 994–1007, 2012.