Proceedings of the GSB Workshop

"Big-data Analysis, IoT and Bioinformatics"

Basic Information

Date	July 11 (Tue), 2017
Venue	Conference Room 11-17, IST, Hokkaido University
Program Committee	Hiroshi HIRATA, Chair (IST, Hokkaido University) Yoshikazu MIYANAGA (IST, Hokkaido University) Hidemi WATANABE (IST, Hokkaido University) Naoki OSADA (IST, Hokkaido University)

List of Contributed Papers

Code	Authors († First author) / Affiliation	Title
BD-1	† Kousuke FUKUI, Akihisa TOMITA Laboratory of Optical Processing and Networking, IST, Hokkaido University	*Analog quantum error correction on encoded qubits for large scale quantum computers*
*BD-2	† Kenta ISHIHARA, Takahiro OGAWA, Miki HASEYAMA Laboratory of Media Dynamics, IST, Hokkaido University	*Detection of Gastric Cancer Risk from X-ray Images based on Machine Learning*
*BD-3	† Keisuke MAEDA, Sho TAKAHASHI, Takahiro OGAWA, Miki HASEYAMA Laboratory of Media Dynamics, IST, Hokkaido University	*Deterioration Level Estimation on Transmission Towers based on Machine Learning*
*BD-4	† Ren TOGO, Kenta ISHIHARA, Takahiro OGAWA, Miki HASEYAMA Laboratory of Media Dynamics, IST, Hokkaido University	*Estimation of regions related to Helicobacter pylori infection from gastric X-ray images*
BD-5	† Namo PODEE, Yoshinori DOBASHI, Tsuyoshi YAMAMOTO Laboratory of Information Media Environment, IST, Hokkaido University	*GPU Adaptive Path Tracing without Atomic Instruction*
BD-6	† Hongjie ZHAI, Makoto HARAGUCHI Knowledge-Base Laboratory, IST, Hokkaido University	*Guessing Associated Features by Non-negative Tri-Factorization*
IoT-1	† Myat Hsu AUNG, Hiroshi TSUTSUI, Yoshikazu MIYANAGA Laboratory of Information Communication Networks, IST, Hokkaido University	*An Implementation of WiFi Based Indoor Positioning System Using Estimated Reference Locations*
IoT-2	† Itaru HIDA,[1] Shinya TAKAMAEDA-YAMAZAKI,[2] Masayuki IKEBE,[1] Masato MOTOMURA,[2] Tetsuya ASAI[1] [1]Laboratoty for Integrated NanoSystem, IST, Hokkaido University; [2]Laboratory for Integrated Digital System Architecture, IST, Hokkaido University	*A Versatile and Energy-Efficient Reconfigurable Accelerator for Embedded Microprocessors*

* These articles are not included in this proceedings due to copyright reasons.

Code	Authors († First author) / Affiliation	Title
IoT-3	† Kodai UEYOSHI,[1] Masayuki IKEBE,[2] Tetsuya ASAI,[2] Shinya TAKAMAEDA-YAMAZAKI,[1] Masato MOTOMURA[1] [1]Laboratory for Integrated Digital System Architecture, IST, Hokkaido University; [2]Laboratory for Integrated NanoSystem, IST, Hokkaido University	*Hardware Accelerator Design for Convolutional Neural Networks with low bit precision*
IoT-4	† Xiaoxiong XING, Yoshinori DOBASHI, Tsuyoshi YAMAMOTO Laboratory of Information Media Environment, IST, Hokkaido University	*Learning interior design using convolutional neural networks*
IoT-5	† Kasho YAMAMOTO,[1] Shinya TAKAMAEDA-YAMAZAKI,[1] Masayuki IKEBE,[2] Tetsuya ASAI,[2] Masato MOTOMURA[1] [1]Laboratory for Integrated Digital System Architecture, IST, Hokkaido University; [2]Laboratory for Integrated NanoSystem, IST, Hokkaido University	*Time-Division Multiplexing Ising Machine on FPGAs*
Bio-1	† Keito AOKI, Kanako KOYANAGI, Hidemi WATANABE Laboratory of Genome Sciences, IST	*Error detection and classifying mixed genomes methods for next generation sequencing based on the characteristics of reads orientation*
Bio-2	† Sangeetha RATNAYAKE, Toshinori ENDO, Naoki OSADA Information Biology Laboratory, IST, Hokkaido University	*Amino Acid Exchangeability and Disease-causing Ability in Human Beta Globin Gene*
Bio-3	† Dai WATABE,[1] Naoki OSADA,[1] Toshinori ENDO,[1] Hiroshi YUASA[2] [1]Information Biology Laboratory, IST, Hokkaido University; [2]The Research Institute of Evolutionary Biology	*American Traditional Bottle Gourds Possessed Hybrid DNA in the Nucleus and Chloroplasts: Alternative Scenario for Ancient Propagation of Lagenaria siceraria*

Analog quantum error correction on encoded qubits for large scale quantum computers

Kosuke Fukui
Graduate School of Information
Science and Technology, Hokkaido University
Kita14-Nishi9, Kita-ku,
Sapporo 060-0814, Japan

Akihisa Tomita
Graduate School of Information
Science and Technology, Hokkaido University
Kita14-Nishi9, Kita-ku,
Sapporo 060-0814, Japan

***Abstract*—Today, big data analytics is valuable for text analytics, machine learning, predictive analytics, data mining, statistics, and businesses. Quantum computation (QC) is an attractive tool to perform faster processing speed for big data analytics, because QC has been shown to solve efficiently some hard problems for conventional computers. Currently, a small scale QC with various quantum systems has been demonstrated. However, a practical QC is still a significant experimental challenge, because of error accumulation. In this work, we propose a method which can alleviate the requirement on error correction for encoded qubits. This novel method improves the tolerance against errors and will pave the way for constructing a practical QCs.**

I. Introduction

Quantum computation (QC) has a great deal of potential [1]. Although small-scale quantum circuits with various qubits have been demonstrated [2], [3], a large-scale quantum circuit that requires scalable entangled states is still a significant experimental challenge for most candidates of qubits. In continuous variable (CV) QC, squeezed vacuum states (SVs) with the optical setting have shown great potential to generate scalable entangled states because the entanglement is generated by only beam splitter (BS) coupling between two SVs [4]. However, scalable computation with SVs has been shown to be difficult to achieve because of the accumulation of errors during the QC process [5]. Because noise accumulation originates from the continuous, nature of the CVQC, it can be circumvented by encoding CVs into digitized variables using an appropriate code, such as Gottesman–Kitaev–Preskill (GKP) code [6], which are referred to as GKP qubits in this work. Menicucci showed that CV-FTQC is possible within the framework of measurement-based QC using SVs with GKP qubits [5]. Moreover, GKP qubits keep the advantage of SVs on optical implementation that they can be entangled by only BS coupling. Hence, GKP qubits offer a promising element for the implementation of CV-FTQC.

To be practical, the squeezing level required for FTQC should be experimentally achievable. Unfortunately, Menicucci's scheme still requires a 14.8 dB squeezing level to achieve the FT threshold $p_{\mathrm{FT}} = 2 \times 10^{-2}$ [7], [8]. Thus, another twist is necessary to reduce the required squeezing level. It is analog information contained in the GKP qubit that has been overlooked because the GKP qubit has been treated as only a discrete variable (DV) qubit. The effect of noise on CV states is observed as a deviation in an analog measurement outcome, which includes beneficial information for quantum error correction (QEC). Harnessing the wasted information for the QEC will improve the error tolerance compared with using the conventional method based on only bit information. In this work, we propose a maximum-likelihood method (MLM) using the analog outcome to improve the error tolerance.

II. Likelihood-function

In this section we review the GKP qubit and likelihood function The GKP qubit, which encode a qubit in an oscillator's q (position) and p (momentum) quadratures, is proposed by Gottesman, Kitaev, and Preskill to correct errors caused by a small deviation in the q and p quadratures [6]. The basis of the GKP qubit is composed of a series of Gaussian peaks of width σ and separation $\sqrt{\pi}$ embedded in a larger Gaussian envelope of width $1/\sigma$. In the case of finite squeezing, the GKP qubits are not orthogonal and there is a probability of misidentifying $|\widetilde{0}\rangle$ as $|\widetilde{1}\rangle$, and vice versa. The probability p_{corr} that we identify the correct bit value is the portion of a normalized Gaussian of a variance σ^2 that lies between $-\sqrt{\pi}/2$ and $\sqrt{\pi}/2$ [5]:

$$p_{\mathrm{corr}} = \int_{\frac{-\sqrt{\pi}}{2}}^{\frac{\sqrt{\pi}}{2}} dx \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-x^2/2\sigma^2). \tag{1}$$

In the measurement we make a decision on the bit value $k(=0,1)$ from the measurement outcome $q_{\mathrm{m}} = q_k + \Delta_{\mathrm{m}}$ to minimize the deviation $|\Delta_{\mathrm{m}}|$, where $q_k(k=0,1)$ is defined as $(2t+k)\sqrt{\pi}(t=0,\pm1,\pm2,\cdots)$, shown in Fig.1(a). If we consider only digital information k, as in conventional QEC, we waste the analog information contained in Δ_{m}. In our method, we propose a MLM to improve our decision for the QEC using analog information. We define the true deviation $|\bar{\Delta}|$ as the difference between the measurement outcome and true peak value $\bar{q}_k$, that is, $|\bar{\Delta}| = |\bar{q}_k - q_m|$. We consider the following two possible events: one is the correct decision, where the true deviation value $|\bar{\Delta}|$ is less than $\sqrt{\pi}/2$ and equals to $|\Delta_{\mathrm{m}}|$ as shown in Fig.1(b). The other is the incorrect decision, where $|\bar{\Delta}|$ is greater than $\sqrt{\pi}/2$ and satisfies $|\bar{\Delta}| + |\Delta_m| = \sqrt{\pi}$, as shown in Fig.1(c). Because the true deviation value obeys

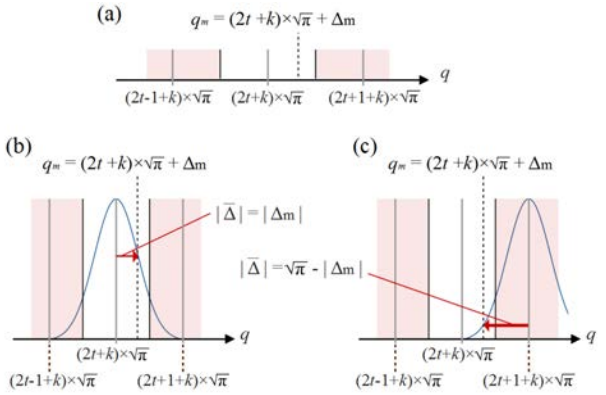


Fig. 1. Introduction of a likelihood function. (a) Measurement outcome and deviation from the peak value in q quadrature. The dotted line shows the measurement outcome $q_{\rm m}$ equal to $(2t+k)\sqrt{\pi}+\Delta_{\rm m}$ $(t=0,\pm1,\pm2,\cdots,\ k=0,1)$, where k is defined as the bit value that minimizes the deviation $\Delta_{\rm m}$. The red areas indicate the area that yields code word $(k+1)$ mod 2, whereas the white area denotes the area that yields the codeword k. (b) and (c) Gaussian distribution functions as likelihood functions of the true deviation value $\bar{\Delta}$ represented by the arrows. (b) refers to the case of the correct decision, where the amplitude of the true deviation value is $|\bar{\Delta}| < \sqrt{\pi}/2$, whereas (c) the case of the incorrect decision $\sqrt{\pi}/2 < |\bar{\Delta}| < \sqrt{\pi}$.

the Gaussian distribution function $f(\bar{\Delta})$, we can evaluate the probabilities of the two events by

$$f(\bar{\Delta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\bar{\Delta}^2/(2\sigma^2)}. \tag{2}$$

In our method, we regard function $f(\bar{\Delta})$ as a likelihood function. The likelihood of the correct decision is calculated by $f(\bar{\Delta}) = f(\Delta_{\rm m})$. The likelihood of the incorrect decision, whose $|\bar{\Delta}|$ is $\sqrt{\pi} - |\Delta_{\rm m}|$, is calculated by $f(\bar{\Delta}) = f(\sqrt{\pi} - |\Delta_{\rm m}|)$. Using this likelihood function, we can reduce the decision error on the entire code word by considering the likelihood of the joint event and choosing the most likely candidate in QEC.

III. Quantum error correction with analog information

We demonstrate that the proposed MLM improves the error tolerance on a concatenated code, which is indispensable for achieving FTQC. The use of a MLM for a concatenated code was proposed with a message-passing algorithm by Poulin [9], and later Goto and Uchikawa [10] for Knill's C_4/C_6 code [7]. Because previous proposals have been based on the probability of the correct decision given by Eq. (1), the error correction provides a suboptimal performance against the Gaussian quantum channel (GQC) [6], [11]. By contrast, our method harnesses analog information using the likelihood functions of measured deviation $f(\Delta_{\rm m})$ and $f(\sqrt{\pi} - |\Delta_{\rm m}|)$. We applied our method to the C_4/C_6 code and simulated the QEC process for the C_4/C_6 code with the conventional [10] and proposed MLM using the Monte Carlo method. In Fig.2, the failure probabilities up to level-5 of the concatenation are plotted as a function of the data qubit's deviation. The results confirm that our method suppresses errors more effectively than the conventional method. It is also remarkable that our method achieves the hashing bound of the standard deviation for the

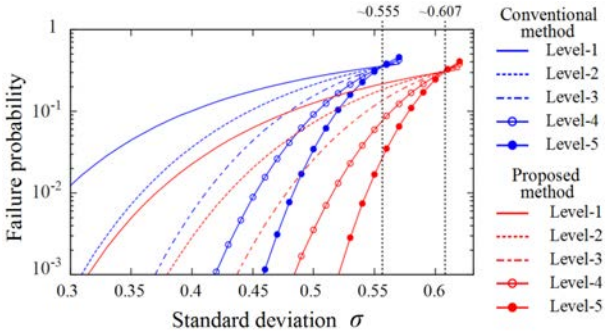


Fig. 2. Simulation results for the failure probabilities of the C_4/C_6 code using the conventional and proposed method. The failure probabilities using the conventional method (blue line) and proposed method (red line) are represented for the concatenated level-1 (solid), level-2 (dashed), level-3 (dashed-dotted), level-4 (open circles), and level-5 (filled circles).

quantum capacity of the GQC $\sim$ 0.607, which corresponds to the squeezing level of 1.3 dB and has been conjectured to be an attainable value using the optimal method [6], [11]. This result implies that our technique improves the GKP qubit into one of the optimal encoded states against the disturbance in the GQC. By contrast, the concatenated code with only digital information achieves the hashing bound $\sim$ 0.555 [6], [11], which corresponds to the squeezing level of 2.1 dB.

IV. Conclusion

We proposed a MLM, which used not only digital information but also analog information for an efficient QEC based on GKP qubits. Numerical results showed that our method improved the error tolerance the concatenated code and is one of the optimal codes against the disturbance of the GQC. Our method will initiate a new approach to build a practical FTQC using both digital and analog information for the QEC. Although it is still difficult to experimentally generate GKP qubits with the squeezing level required for FTQC [5], our method can alleviate this requirement and will encourage experimental developments.

Acknowledgment

This work was funded by ImPACT Program of Council for Science, Technology and Innovation.

References

[1] P. Shor, In Proceeding of 35th IEEE FOCS, pp.124-134, Santa Fe, NM, Nov 20-22 (1994).
[2] T. Niemczyk, F. Deppe, H. Huebl, E. P. Menzel, F. Hocke, M. J. Schwarz, J. J. Garcia-Ripoll, D. Zueco, T. Hümmer, E. Solano, A. Marx, and R. Gross, Nat. Phys. **6**, 772-776 (2010).
[3] R. Blatt and D. Wineland, Nature **453**, 1008-1015 (2008).
[4] J. Yoshikawa, S. Yokoyama, T. Kaji, C. Sornphiphatphong, Y. Shiozawa, K. Makino, and A. Furusawa, APLPhotonics **1** 060801 (2016).
[5] N. C. Menicucci, Phys. Rev. Lett. **112**, 120504 (2014).
[6] D. Gottesman, A. Kitaev, and J. Preskill, Phys. Rev. A **64**, 012310 (2001).
[7] E. Knill, Nature, **434**, 39-44 (2005).
[8] K. Fujii and K. Yamamoto, Phys. Rev. A **82**, 060301 (2010).
[9] D. Poulin, Phys. Rev. A **74**, 052333 (2006).
[10] H. Goto and H. Uchikawa, Sci. Rep. **3**, 2044 (2013).
[11] J. Harrington and John Preskill, Phys. Rev. A **64**, 062301 (2001).

GPU Adaptive Path Tracing without Atomic Instruction

Namo Podee
Hokkaido University
Sapporo, Japan
Email: namo@ime.ist.hokudai.ac.jp

Yoshinori Dobashi
Hokkaido University
Sapporo, Japan
Email: doba@ime.ist.hokudai.ac.jp

Tsuyoshi Yamamoto
Hokkaido University
Sapporo, Japan
Email: yamamoto@ist.hokudai.ac.jp

***Abstract*—We present an adaptive technique for path tracing on a GPU without the use of atomic instruction. The technique improves the efficiency of the current state of the art parallel path tracing methods. Our method uses a stream compaction algorithm to generate, in parallel, a list of pixels to be traced, also called a sample stream, which may contain multiple samples for each pixel. To accelerate the convergence, we choose pixels to be traced by predicting the square error reduction rate, which is computed by comparing the past path tracing result and its filtered version with a bilateral filter. Then, we use traditional stream compaction path tracing for the generated sample stream and accumulate the result iteratively, in parallel. We show that our method is up to 2.6 times faster compared to previous parallel path tracing techniques for equal-quality rendering. We also analyze how much improvement has been achieved in different scenes and discuss the limitations of our method.**

I. Introduction

Light transport problems used to be practically impossible to solve interactively because of their complexity. However, the advances made in general purpose graphics processing units (GPGPU), now means these problems can be solved interactively by using a massive amount of threads on a GPU.

While a naive GPU implementation can speed up the path tracing algorithm, this does not take full advantage of the computing power of the GPU due to the uneven workload between threads. To solve this, several methods [1] [2] [3] [4] that make better use of the parallel computing capability have been implemented. However, little research has been done on efficient adaptive path tracing on a GPU.

In this paper, we propose a solution to this problem; we propose adaptive path tracing on a GPU with minimal synchronization. Our method spends the computing power on the erroneous pixels, instead of the whole image, and achieves improvements in the overall path tracing performance. From our experiments, we conclude that our method can increase the computation speed by up to a factor of 2.6 compared to previous parallel path tracing methods with equal-quality rendering.

II. Adaptive Path Tracing without Atomic Instruction

There are two main problems in adaptive path tracing on a GPU: thread-pixel assignment and the accumulation of results. The assignment problem is how to assign a pixel to each thread. It is complicated to efficiently assign pixels to each thread because we have to do everything in parallel, implying that we do not have synchronization between threads, since the parallel adaptive method usually processes one pixel on multiple threads; thus a naive adaptive implementation uses atomic operations to assign samples to threads. Next, the accumulation problem is in regard to the output from adaptive path tracing; it is most likely that we get multiple results for the same pixel because we try to trace erroneous pixels multiple times. With these results for the same pixel, we should not naively write the result to the image buffer simultaneously because a memory access conflict will occur and this slows down the computation.

We have developed a technique that can efficiently generate a list of pixels to be traced, which we call a sample stream. This sample stream is proportional to the estimated error reduction ratio for each pixel and the result can still be efficiently accumulated from a path tracer to the image buffer in parallel without causing conflict.

A. Method Overview

Figure 1 illustrates our method. Our method begins with creating an initial image by tracing a set of paths for each pixel. The number of paths is the same for all pixels. We use the light vertex cache bidirectional path tracing method developed by Davidovic et al. [1] because it is the most efficient GPU path tracing method we have tested. The image obtained by this process is then filtered with a bilateral filter. We calculate the difference between the unfiltered and filtered intensities of the pixels then use this to estimate the error reduction, which tells us how much the difference between the final result and the current result will be reduced by after some additional samples. After that, we use the error reduction data as a probability to determine whether additional paths from the viewpoint (eye paths) should be traced for each pixel or not. Next, we generate multiple sets of pixels to be traced iteratively based on the probability.and combine all these sets together and call this a sample stream. We then trace the eye paths for the pixels in the sample stream and accumulate the contributions from each set in the image buffer.

III. Result

Figure 2 shows the root mean square error (RMSE) of the result using our method, compared with the stream compaction

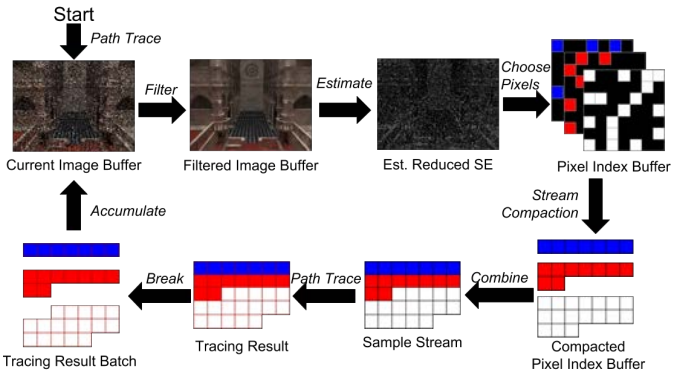


Fig. 1. An overview of our method.

method [5] and the adaptive atomic instruction method, plotted against rendering time for three scenes. The adaptive atomic instruction method was implemented by initially sampling a scene, filtering it with the bilateral filter then calculating the square error value of each pixel just as in our method.

the result via atomic instructions on a GPU. As demonstrated by Figure 2, the convergence speed using our method is faster than those using the other methods. For example, when compared to the methods using stream compaction and atomic instructions only, our method was up to 2.6 and 3.0 times faster, respectively. The method using atomic instruction underperformed when the system could not approximate the error of each pixel accurately enough. We should be able to improve our method efficiency by adjusting the sample stream layout, but due to the time limitation, we plan to adjust it in future work.

IV. Conclusion

We have proposed an efficient adaptive path tracing method on a GPU. Our method outperforms the previous state of the art GPU path tracing techniques in that it is up to 2.6 times faster for equal-quality rendering. The outperformance continues as long as the filtered image can faithfully estimate the square errors of the result. However, due to the time limitation we cannot compare our method against other novel adaptive path tracing methods, which we plan to do in future work. Our filter may fail to faithfully estimate the square errors when we use parameters inappropriate for the scene to be rendered. Our filter requires many parameter adjustments, which are not intuitive and may be a tedious task. It is interesting to solve this problem by using a machine learning approach or more advanced filters. Our method can also be improved by adjusting the sample stream layout to be more cache friendly and thereby increase its performance.

Acknowledgment

Cornell Box model courtesy of Cornell University, Conference Room model courtesy of Anat Grynberg and Greg Ward, Sibenik cathedral model courtesy of Marko Dabrovic.

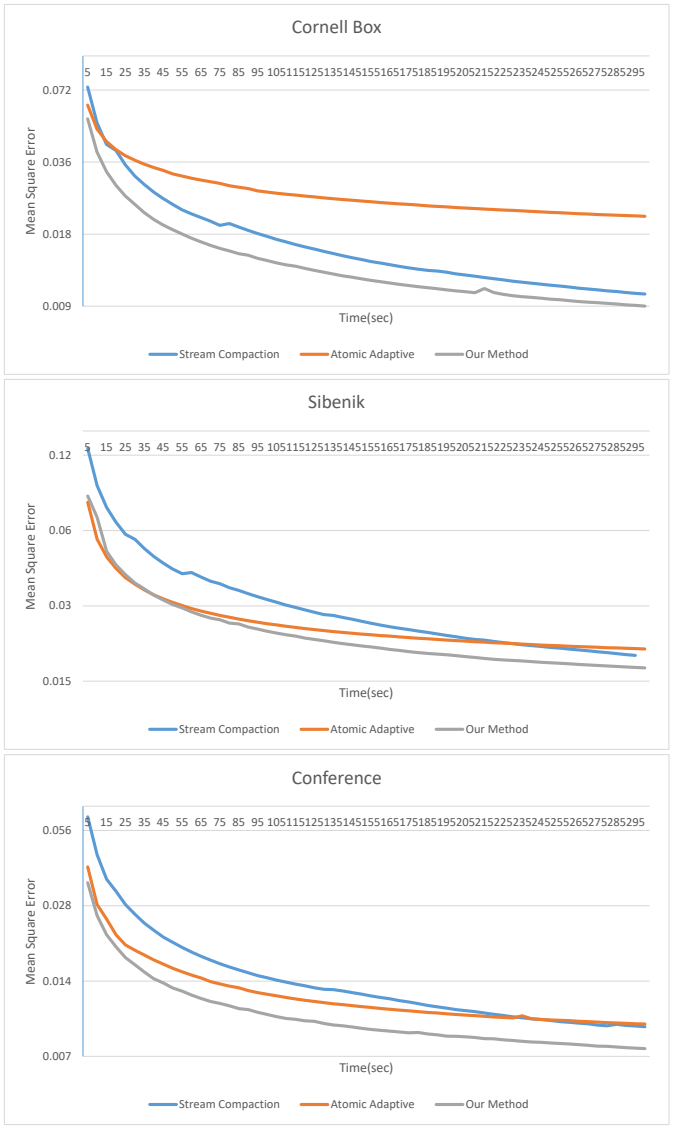


Fig. 2. Convergence-time plot of the three scenes in logarithmic scale.

References

[1] T. Davidovič, J. Křivánek, M. Hašan, and P. Slusallek, "Progressive light transport simulation on the gpu: Survey and improvements," *ACM Trans. Graph.*, vol. 33, no. 3, pp. 29:1–29:19, Jun. 2014. [Online]. Available: http://doi.acm.org/10.1145/2602144

[2] J. Novk, V. Havran, and C. Dachsbacher, "Path Regeneration for Interactive Path Tracing," in *Eurographics 2010 - Short Papers*, H. P. A. Lensch and S. Seipel, Eds. The Eurographics Association, 2010.

[3] I. Wald, "Active thread compaction for gpu path tracing," in *Proceedings of the ACM SIGGRAPH Symposium on High Performance Graphics*, ser. HPG '11. New York, NY, USA: ACM, 2011, pp. 51–58. [Online]. Available: http://doi.acm.org/10.1145/2018323.2018331

[4] D. van Antwerpen, "Improving simd efficiency for parallel monte carlo light transport on the gpu," in *Proceedings of the ACM SIGGRAPH Symposium on High Performance Graphics*, ser. HPG '11. New York, NY, USA: ACM, 2011, pp. 41–50. [Online]. Available: http://doi.acm.org/10.1145/2018323.2018330

[5] M. Billeter, O. Olsson, and U. Assarsson, "Efficient stream compaction on wide simd many-core architectures," in *Proceedings of the Conference on High Performance Graphics 2009*, ser. HPG '09. New York, NY, USA: ACM, 2009, pp. 159–166. [Online]. Available: http://doi.acm.org/10.1145/1572769.1572795

Guessing Associated Features by Non-negative Tri-Factorization

Hongjie Zhai, Haraguchi Makoto
Graduate School of Information Science and Technology, Hokkaido University
Kita 14, Nishi 9, Kita-ku, Sapporo, Hokkaido, 060-0814, Japan
Email: zhaihj@kb.ist.hokudai.ac.jp

***Abstract*—In this paper, we try to guess feature associations from limited knowledge. That is, given two object sets with their own feature sets, our task is to guess associations between features, where only a small part of associations is presented. We call the known associations as "hints". To achieve this goal, we build common clusters cross all the features, where the known associated features will be clustered into a common cluster. For other features, they will be clustered based on their similarities with hints. Technically speaking, we use the Non-negative Tri-factorization to do the clustering on all features. A laplacian constraint is proposed to guarantee associated features will be put in the same cluster. We experimentally show that the proposed method can guess many meaningful associations.**

I. Introduction

In this paper, we study a problem of guessing feature association. That is, considering two object sets $O^1 = \{o_1^1, o_2^1, \ldots, o_n^1\}$ and $O^2 = \{o_1^2, o_2^2, \ldots, o_m^2\}$, where objects in O^1 are described by feature set $F^1 = \{f_1^1, f_2^1, \ldots, f_u^1\}$ and objects in O^2 are described by another feature set $F^2 = \{f_1^2, f_2^2, \ldots, f_v^2\}$. Some "hints" of associations, which are partial associations between two feature sets, are assumed already known. The target to finding the association between remaining features.

We proposed a novel method for finding the feature associations. To guess the missing associations, our method tries to find new associations by the knowledge from known associations (hints). Once new associations are detected, they can be added to known association set. Thus, we are able to use the new known association set to furtherly find new associations. By repeating this process, finally all the possible associated feature pairs can be detected. For the performance and scalability, we formulize this clustering-based methed by Non-negative Tri-factorization and experimentally show its ability for guessing associations.

II. Basic Idea and Algorithm

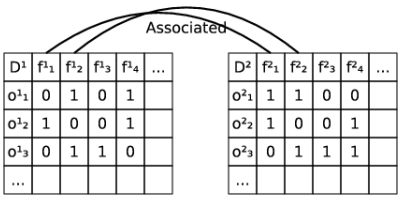
The basic idea of our method is illustrated in Figure 1. Here, we have two object-feature relation tables, where $f_1^1, f_2^1, f_3^1, \ldots$ are the features of objects $o_1^1, o_2^1, o_3^1, \ldots$. If a object contains the feature, the corresponding position will be 1 as shown in Figure 1a. Additionally, we also have two associated feature pairs: $f_1^1 - f_1^2$ and $f_2^1 - f_2^2$. Firstly, we merge the associated features into one. After that, by only focus on the merged ones, we can construct the vector representation of objects with the same dimensions. For example, in Figure 1b, f_1^1 and f_1^2 are merged as f_1 as well as that f_1^2 and f_2^2 are merged into f_2. With these constructed vectors, we perform clustering on objects. As the result, the objects in different sets may be clustered into one object cluster just like Figure 1b. Here, cluster c_1 contains object o_1^1, o_3^1 and o_3^2 while cluster c_2 contains object o_2^1 and o_2^2. Moreover, by representing features with object clusters, we can perform clustering on features to find new associations. This is illustrated in Figure 1c. It is easy to find that under the object cluster representation, f_4^1 and f_4^2 have the same vector. Thus, we found a new association $f_4^1 - f_4^2$. Once new associations are found, we merge into one and repeat the whole process until no new associations can be found.

To improve the performance for large scale data, instead of the two-phases algorithm, we embed features into a common space by tri-factorization proposed by [Long 05]. In this common space, the associated features are guarantee to have the same vector representation. We formulated our idea into the algorithm 1.

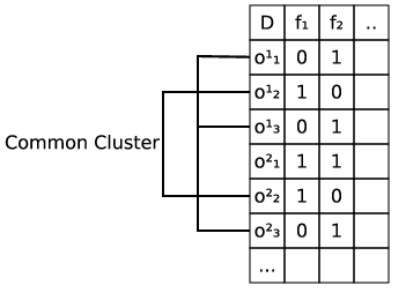
Data: Object-feature relation matrix: D^1, D^2, feature relation matrix: W, feature set F^1 and F^2, object set O^1, O^2, Dimension Parameter: N,M
Result: Associated feature pairs
Initialize random non-negative matrix L^1, C^1, R^1, L^2, C^2, R^2, where $C^{\{1,2\}} \in \mathbb{R}+^{N \times M}$;
$K = D - W$, where $d_{ii} = \sum_j w_{ij}$;
Solve the following tri-factorization problem:
$argmin_{L^1,C^1,R^1,L^2,C^2,R^2} |D^1 - L^1C^1R^1| + |D^2 - L^2C^2R^2| + \lambda(\frac{R^1}{R^2})^T K(\frac{R^1}{R^2})$;
Assign column vector in R^1 and R^2 to each feature in order;
for *each vector* v_j^1 *in* R^1 **do**
  Find the nearest vector v_j^2 in R^2;
  Print (F_i^1, F_j^2) as associated feature pair.
end
Algorithm 1: Tri-factorization for Association Learning

Here, $D^1 \in \mathbb{R}+^{|O^1| \times |F^1|}$, $D^2 \in \mathbb{R}+^{|O^2| \times |F^2|}$ are the relation matrix, where $d_{ij} = 1$ if object o_i contains feature f_j. $W \in \mathbb{R}^{|F^1|+|F^2| \times |F^1|+|F^2|}$ is called feature relation matrix. It is constructed from the following rules:

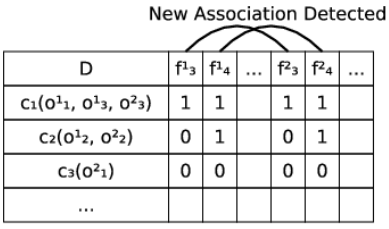
- if $i \leq |F^1| and j \leq |F^1|$, w_{ij} = 0
- if $i \geq |F^1| and j \geq |F^1|$, w_{ij} = 0



(a) Merge associated features



(b) Clustering in common feature space



(c) Mining new associations in common object cluster space

Fig. 1: Illustration of Idea

- if $i \leq |F^1| and j \geq |F^1|$, if f_i and f_j are known to be associated $w_{ij} = 1$, else $w_{ij} = 0$
- if $i \geq |F^1| and j \leq |F^1|$, $w_{ij} = w_{ji}$

By considering each feature as a vertex and connect the associated feature pairs, we can get a bi-graph G. It is easy to know that the matrix K is the laplacian matrix of graph G. According to [Cai 11], the laplacian constraint can make sure the associated features always have similar vector representation in the common space.

III. Experiment

To validate the ability of proposed method, we performed a preliminary experiment. We take the english/french news articles between 1996-08-20 and 1996-08-25 from Reuters Corpora [Lewis 04]. Detailedly speaking, the english news articles are selected from RCV1 (i.e. Reuters Corpus Volume 1) while the french news articles are selected from RCV2 (i.e. Reuters Corpus Volume 2). Articles are the objects and words are treated as features. We use all the $2,000$ articles in french and randomly sampled $2,000$ english articles from total about $20,000$ articles to balance the size of dataset. After morphological analysis by tree-tagger [Schmid 13], we only keep nouns. As the result, the number of french words is $4,783$ and for english, it is $6,020$. By using english-french dictionary, we select 500 pairs of word with similar meaning as the associated features. By applying the proposed algorithm, we get the vector representation of words in a common space. We set λ to 300, N to 250 and M to 250. After this, to show the ability of finding associations, we clustered english/french into common clusters. If our algorithm works, we should find associated words in the common cluster, excluding the given associated words. Because of the space limitation, here we only give one example of the common clusters. As shown in table I, for the ease of reading, the common cluster are splitted into english words and french words. We can see that many associated words are successfully clustered into the common cluster, e.g. cardiologue and cardiologist, peau and skin, etc.

French	English
température malaria	*temperature* malaria
cardiologue accordéon	cardiologist rigor
peau respirateur	skin respirator
élu degustation	chill pneumonia
ombre bijou	saint FSA
bras jambe	bouquet heartbeat
occurrence idylle	pacemaker investor
...	...

TABLE I: Content of common cluster

IV. Conclusion and Future Works

This paper proposed a general method for feature association guessing and give a tri-factorization formulation of the method. The tri-factorization formulation enjoys both the performance from NMF and the generalness of our idea. We also showed that the algorithm works as expected through a preliminary experiment. However, because of the deviation of data, we failed to get the "common sense" associations (i.e. the associated words in dictionary). In future, to solve this problem, (1) we would like to construct a more consist dataset. For example, we can select in a longer duration as well as limit the categories. (2) Because many words have different meanings under different contexts, we also would like to take semantic information into our consideration.

References

[Lewis 04] Lewis David D., et al. "Rcv1: A new benchmark collection for text categorization research." Journal of machine learning research 5.Apr (2004): 361-397.

[Cai 11] Cai Deng, et al. "Graph regularized nonnegative matrix factorization for data representation." IEEE Transactions on Pattern Analysis and Machine Intelligence 33.8 (2011): 1548-1560.

[Ding 06] Ding Chris, et al. "Orthogonal nonnegative matrix t-factorizations for clustering." Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006.

[Long 05] Long Bo, Zhongfei Mark Zhang and Philip S. Yu. "Co-clustering by block value decomposition." Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. ACM, 2005.

[Chakraborty 15] Chakraborty Yulong Pei1 Nilanjan and Katia Sycara. "Nonnegative matrix tri-factorization with graph regularization for community detection in social networks." (2015).

[Lee 01] Lee Daniel D. and H. Sebastian Seung. "Algorithms for nonnegative matrix factorization." Advances in neural information processing systems. 2001.

[Schmid 13] Schmid, Helmut. "Probabilistic part-ospeech tagging using decision trees." New methods in language processing. Routledge, 2013.

An Implementation of WiFi Based Indoor Positioning System Using Estimated Reference Locations

Myat Hsu Aung, Hiroshi Tsutsui, and Yoshikazu Miyanaga
Graduate School of Information Science and Technology, Hokkaido University
Kita 14, Nishi 9, Kita-ku, Sapporo, Hokkaido 060-0814, Japan

***Abstract*—In this paper, we present an implementation of WiFi-based indoor positioning system using estimated reference locations. In case of general WiFi based indoor positioning systems, the database of WiFi access points is constructed by gathering pairs of MAC address and received signal strength indicator (RSSI) value of each known reference location. However, this task requires high cost since the administrator should know the actual position of each reference location. In the proposed approach, the database is constructed by gathering MAC-RSSI pairs using a reference device moving in a constant speed with simple direction. Assuming a constant speed, the location of each reference point can be estimated from the velocity. Estimation accuracy evaluation results show that user's locations can be roughly estimated.**

I. Introduction

Recently, positioning systems have been widely used in consumer devices. Users' current position information may help users to navigate themselves to their destination in outdoor environment and even in indoor environment. For indoor positioning systems (IPS), RF signals from wireless local area network (WLAN), Bluetooth, and cellular networks can be utilized to estimate the users' positions [1]. Among them, we focus on WiFi based IPS since the WiFi coverage is getting higher due to greatly increasing number of private or public WiFi access points in metropolitan areas.

There are a lot of issues to be tackled in IPS. In case of fingerprinting based IPS, the information of access points reachable from user's location is used for estimating the user's location by comparing the pre-stored data in the database. Such user's information is called fingerprint. As for database, we need store the information of access points reachable from a lot of reference points whose locations are known. This database creation requires quite large cost since many data should be collected with their actual positions using precise indoor map or floor plan.

Motivated by this, we are proposing a WiFi based IPS using estimated reference locations [2]. We are trying to develop a method to create database of reference point information which does not require precise reference point locations. Using the proposed system, it is expected the cost of database creation can be dramatically reduced. In the proposed approach, the database is constructed by gathering MAC-RSSI pairs using a reference device moving in a constant speed with simple direction. Assuming a constant speed, the location of each reference point can be estimated from the velocity. In this paper, we present the implementation of the proposed approach with an accuracy evaluation. Evaluation results show that user's locations can be roughly estimated.

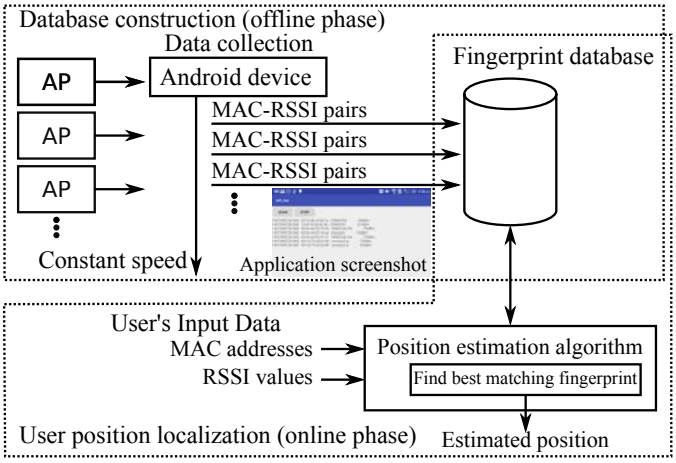


Fig. 1: The overview of the proposed system.

II. Proposed approach

Received signal strength indicator (RSSI) is one of the most widely used cues for indoor localization. The RSSI values from wireless access points can be used to estimate the position of user's mobile device. In general indoor localization systems, a user sends data of wireless access points reachable from the current position to the server and receives the location information.

There are several approaches to estimate positions such as trilateration/triangulation and fingerprinting [3]. In this paper, we focus on the fingerprinting approach since this approach does not require additional hardware installation such as dedicated beacon devices. In this approach, only RSSI values from available access points are required. Also the actual location of each access point does not required. The user positions are estimated by finding the best match of the real time RSSI values in RSSI fingerprint database.

Figure 1 shows the overview of the proposed system. The total system consists of a database construction part (offline phase) and a user position localization part (online phase).

In the database construction part, pairs of MAC address and RSSI value which can be obtained from reachable access points (APs) from each reference point are collected in the target area by using such as WiFi scanning Android OS applications. These MAC-RSSI pairs are gathered every specific period such as one second with walking at a constant speed to estimate the actual location of each reference point. The

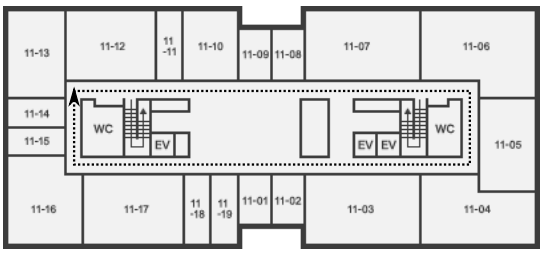
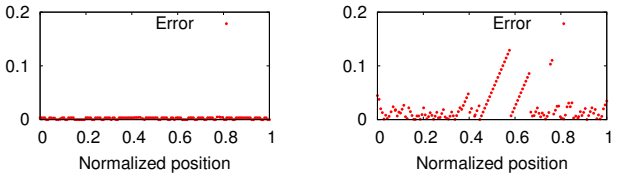
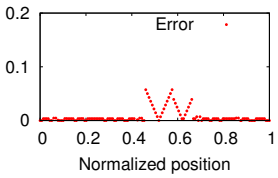


Fig. 2: Floor map and direction in the experiment. Normalized position 0 corresponds to the starting point and 1 to the end point.

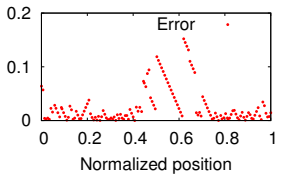


(a) DB: D1, input: D1.

(b) DB: D1, input: D2.



(c) DB: D2, input: D2.



(d) DB: D2, input: D1.

Fig. 3: Normalized error of each estimated position.

absolute timestamps for each pair is also collected. As a result, a set of pairs of MAC addresses and RSSI values for each reference point is stored as fingerprints in the database.

In the user position localization part, user's mobile device samples the MAC address and the RSSI value of each access point available at the current user's position. The procedure of the user's position estimation is the following.

(1) Pick up reference point whose MAC address list includes at least one MAC address of user's input, and create the set of such reference points. This set is described by S.
(2) Create the list of unique MAC addresses of user's input data and S. This list is decribed by M.
(3) Create location vector, that is fingerprint, for user's input and references points S. Each element of a location vector is given by RSSI value for correspoing MAC address listed in M. If no RSSI value is avaliable for a MAC address, the element is set $\varnothing$.
(4) Calculate vector distances between user's input vector and all reference vectors.
(5) Output the position which gives minimum distance as the estimated user's position.

III. Implementation and Experimental Results

We implemented a part of our approach as an Android application which gathers MAC-RSSI pairs of available access points every specific period. Data is collected in 11th floor of Graduate School of Information and Science Technology Building, Hokkaido University, whose floor map is shown in Fig. 2. The RSSI values are collected from all available access points by walking on the dashed line about 120 m in Fig. 2 with HTC One (M7) android device at constant speed.

The accuracy of the proposed system is evaluated using the user's input data from the android device. We obtained two sets of data assuming one is for database creation (training) and the other is for position estimation (testing). These are denoted by D1 and D2 in the following. The numbers of samples are slightly different with each other due to slight speed change. Therefore, timestamp normalization is required.

To estimate the user's location, we need to define vector distances $D(V_u, V_i)$ between user's input vector V_u and reference vector V_i. In this experiment, the following distance is used.

$$D(V_u, V_i) = \frac{1}{N^A} \sum_{\{j|j\in M, V_{uj}\neq\varnothing, V_{ij}\neq\varnothing,\}} |V_{uj} - V_{ij}|^2, \quad (1)$$

where $N = \#\{j|j \in M, V_{uj} \neq \varnothing, V_{ij} \neq \varnothing, \}$, $A \geq 1$, and we set 3 to A. This parameter A controls the contribution of N', where larger N gives smaller distance. Also thresholding based on N is utilized to remove noise. If $N < 10$, we set large value to the distance.

Figure 3 shows the errors for the Android device. When the same data set is used for both database creation and user's position estimation, it is obvious that the error is small such as under 0.05 in normalized error, which corresponds to 6 m, as shown in Figs. 3 (a) and (c). In Fig. (c), some large errors can be found. This is because P_i for some neighboring i values are identical. When the different data sets are used for database creation and user's position estimation, larger error occurs compared to the previous case, as shown in Figs. 3 (b) and (d). However, the maximum error is 0.17 in normalized error, which corresponds to 20 m.

IV. Conclusion

In this paper, a WiFi based IPS using estimated reference locations is proposed. In the proposed approach, the fingerprint database is constructed by gathering MAC-RSSI pairs using a reference device moving in a constant speed with simple direction. Estimation accuracy evaluation results show that the proposed approach can estimate user's location with maximum error of 20 m without any precise reference point locations.

References

[1] R. Yasmine and L. Pei, "Indoor fingerprinting algorithm for roon level accuracy with synamic database," in *Proc. Fourth International Conference on Ubiquitous Positioning, Indoor Navigation and Location Based Services (UPINLBS)*, Nov. 2016, pp. 113–121.
[2] M. H. Aung, H. Tsutsui, and Y. Miyanaga, "An accuracy evaluation of WiFi based indoor positioning system using estimated reference locations," in *Proc. SISA 2017*, Sep. 2017, to appear.
[3] W. K. Zegeye, S. B. Amsalu, Y. Astatke, and F. Moazzami, "WiFi RSS fingerprintion indoor localization for mobile devices," in *Proc. IEEE 7th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*, Oct. 2016, pp. 1–6.

A Versatile and Energy-Efficient Reconfrigurable Accelerator for Embedded Microprocessors

Itaru Hida, Shinya Takamaeda-Yamazaki, Masayuki Ikebe, Masato Motomura and Tetsuya Asai
Graduate School of Information Science and Technology, Hokkaido University
E-mail: hida@lalsie.ist.hokudai.ac.jp

***Abstract*— Conventional processors are energy in-efficient in that they fail to utilize the fact that most of their time and energy are spent on heavily-recursively executed small code segments. A DYNaSTA accelerator, proposed and implemented, is an architectural solution to such a problem. Not only exhibiting around an order of magnitude energy efficiency improvement, the architecture can also exploit full potential of the low-power circuit techniques such as DVFS and power gating.**

I. Introduction

Overwhelming trends toward Internet of Things explain why low energy embedded microprocessors (EMPs) are getting more important than ever. Sources of energy inefficiency in EMP architectures are fairly well understood: the needs 1) to fetch/decode every instruction from memory, 2) to write/read register files to acquire/store operands per every instruction, 3) and to clock numerous numbers of F/Fs for pipelining multiple instructions on a datapath [1]. Given this insight, we may choose to "statically" map those instructions in heavily exe-cuted “recursive codes” to array of ALUs prior to their execu-tion. By running the codes just as combinatory datapath with no registers, 1) to 3) redundancies can be drastically reduced. Though this “reconfigurable accelerator” solution looks straightforward and attractive, there is an inherent drawback: it is hard to cope with complex control flows (i.e., lots of branches) typical in embedded applications. It explains why previous such proposals have focused on simple code seg-ments that do not have a branch [2]..

II. Architecture

Based on this observation, we have recently proposed abstract architecture for achieving both energy efficiency and versatility in control-rich embedded applications [3]. Contribution of this paper is to materialize the concept into executable micro-architecture, design/verify it in real silicon chip, and evaluate its energy efficiency. Key innovation in our proposal, a DYNaSTA reconfigurable accelerator shown in Fig. 1, is to combine two distinctive array structures different in nature, namely dynamic operand forwarding matrix (DYN), and static ALU array (STA). STA plays a key role in achieving energy efficiency, while DYN does so in versatility.

STA features non-fixed number of stages, where each stage has several ALUs sharing a set of source/destination lines (Fig. 1). For reducing the number of switches hence improving en-ergy efficiency, only parallel instructions are mapped onto a same stage, where branch/jump and load/store instructions go to the 1st and last ALUs, respectively (Fig. 2(b) and (d)). The instructions dependent on preceding ones are mapped onto the next stage. Conditional execution is supported for discarding short forward branches. Appropriate number of STA stages is dependent on the sizes of target codes, while that of ALUs per stage will range from 2 to 8 as

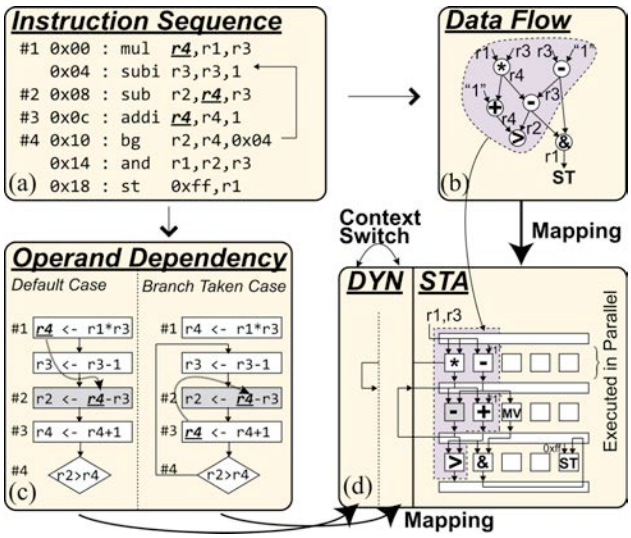


Fig. 2. Code mapping policy: (a) an example code, (b) extracted data flow, (c) extracted operand dependency, (d) mapping on DYN and STA.

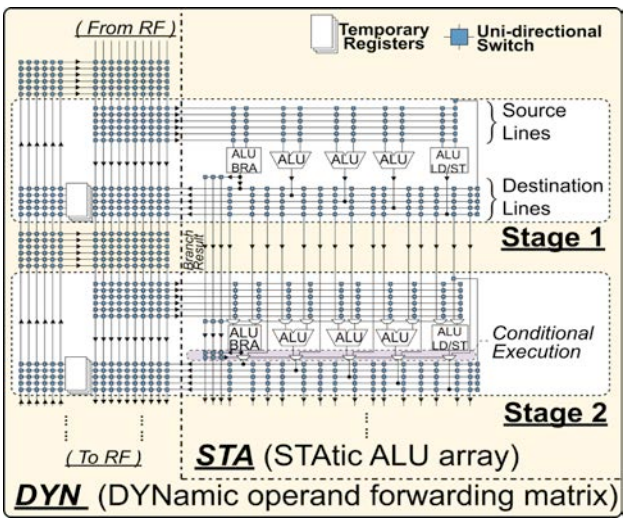


Fig. 1. DYNaSTA reconfigurable accelerator architecture.

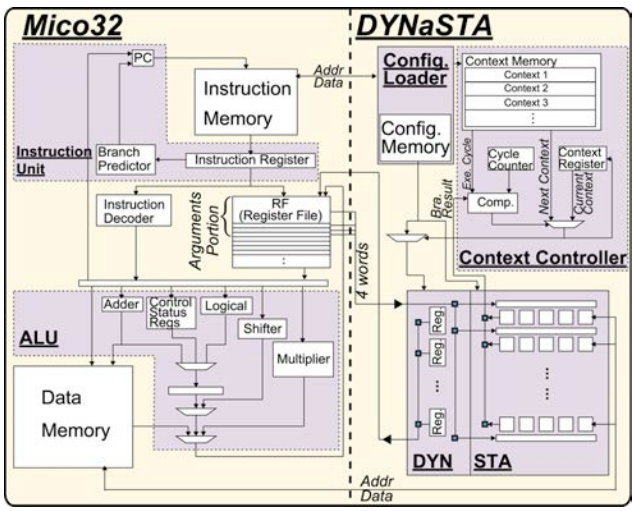


Fig. 3. Tight-integration of Mico32 (base EMP) and the DYNaSTA accelerator.

in superscalar/VLIW architec-tures. Note there is no registers hence no clocks in STA.

Difficulty in serving branches in a reconfigurable accelerator lies in that their outcome can never be known apriori: for example (Fig. 2(c)), the "r4" operand in #2 may be produced by #3 instead of #1 when #4 branch is taken. Efforts to ac-commodate such dynamic nature in the ALU array like STA unavoidably degrade its simplicity and regularity hence in-curring energy inefficiency. DYN is a multi-context, bidirec-tional operand forwarding matrix for solving this difficulty: it is reconfigured only when operand dependencies amongst in-structions are altered on a branch (Fig. 2(c) and (d)). Operands that affected are memorized/forwarded in/from temporary registers. Keeping power-consuming dynamic reconfiguration away from the massive ALU array (and leaving it static) is a key for achieving energy efficiency in DYNaSTA architecture.

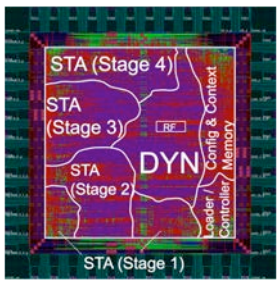
III. Implementation

We have designed an EMP with this DYNaSTA accelerator into silicon (Fig. 3). Base EMP is Mico32 [4], which has been chosen because of its typical RISC architecture and opensource RTL code. The accelerator also includes a config-uration loader, which sets whole DYNaSTA configuration prior to execution, and a context controller, which directs re-configuration of DYN. By treating recursive codes that are mapped onto DYNaSTA as subroutines, the read/write path between Mico32's RF and DYN only needs to cover its ar-guments portion (4 registers, Fig. 3). We have implemented this design with using UMC 0.18um process (Fig. 4 and Table 1). Though the size of DYNaSTA is very small, extending it is quite straightforward. The chip is now under fabrication.

IV. Evaluation

Performance and power consumption of the DYNaSTA accelerator are evaluated using sample applications (Table 2) based on the synthesized netlist.

Fig. 5 compares power consumption of these codes running on Mico32 and the DYNaSTA. 69% to 86% reductions are observed due mainly to discarded instruction memory access. Logic power consumption is also reduced, whose detailed breakdown is shown in Fig. 6. From Figs. 5 and 6, it is clear that 1) to 3) redundancies mentioned earlier have been successfully removed. Since instructions are executed in parallel in STA (Fig. 2(d)), the proposed architecture not only reduces the power but also enhances its performance (Fig. 7) at a same frequency (100MHz). Resultantly, energy efficiency is improved by 4.5 to 13 times from Mico32 for these sample codes.



.Fig. 4. Chip layout view.

TABLE I. Chip Specification.

Specification	
Technology	UMC 0.18um 1P6M CMOS
Package	48-pin DIL
Core Area	1.06 mm x 1.06 mm
Gate Count	86.5K
Supply Voltage	1.8 V core 3.3 V IO
Clock Frequency	100 MHz
# of Stages	4

TABLE II. Summary of Sample Applications.

Application	# of Instrs.	# of Brs.	PE Utilization [%]	# of Contexts
Fibonacci	12	3	24	5
Sbox	25	2	50	5
CRC32	18	2	36	5
Sepia Filter	22	1	44	3

V. Conclusion

Table 3 reveals reasons of this energy efficiency: though DYNaSTA consumes x18.5 more gates than Mico32, its average toggle rate is as low as x0.06 of Mico32. Specifically, gate-consuming STA features only 1.8% toggle rate, which accounts for its relatively low power occupation in Fig. 6.

Filling a chip with simple, regular and energy-efficient array like DYNaSTA can become an interesting solution in "Dark Silicon" [5] era (Fig. 8). Here, existing domain-oriented low power circuit techniques such as DVFS and power gating can augment the architecture quite nicely. For instance, since only a few active stages propagate like a "wave" on the array, re-maining numerous "silent" stages can be powered-off systematically for minimizing leak current (Fig. 8). Our next challenges include enhancing DYNaSTA with such low-power circuit techniques as well as establishing code mapping SW.

References

[1] R. Hammed, et al., "Understanding sources of inefficiency in general purpose chips," ACM Com. Arch., no.38, pp.37-47, 2010.

[2] N. Ozaki, et al., "Cool Mega-Arrays: ultralow-power reconfigu-rable accelerator chips," IEEE Micro, vol.31(6), pp.6-18, 2011.

[3] T. Hirao, et al., "A restricted dynamically reconfigurable architecture for low power processors," ReConFig 2013, 3A-1, 2013.

[4] http://en.wikipedia.org/wiki/LatticeMico32.

[5] H. Esmaeilzadeh, et al., "Dark silicon and the end of multicore scaling," ISCA 2011, pp. 365-376, 2011.

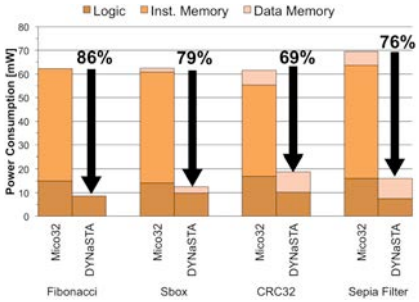


Fig. 5. Mico32 vs. DYNaSTA: total power consumption of sample applications.

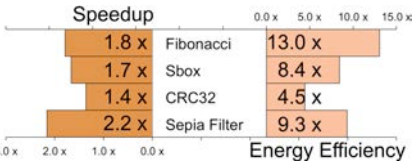


Fig. 7. DYNaSTA/Mico32 performance and energy efficiency improvements.

TABLE III. Mico32 vs. DYNaSTA: Gate Counts and Average Toggle Rates (Fibonacci).

		Gate Count [k Gates]	Average Toggle Rate [%]
DYNaSTA	DYN	43.6	20.0
	STA	322.9	1.8
	Others	80.2	9.1
	All	446.8	4.9
Mico32		24.1	76.8
Ratio (DYNaSTA /Mico32)		18.48	0.06

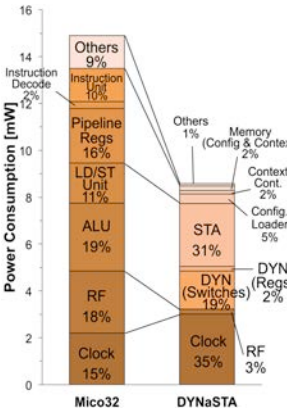


Fig. 6. Mico32 vs. DYNaSTA: logic power consumption (Fibonacci).

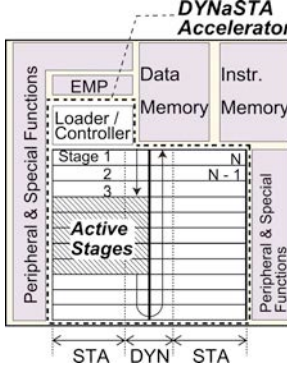


Fig. 8. DYNaSTA SoC concept toward "Dark Silicon" era.

Hardware Accelerator Design for Convolutional Neural Networks with Low Bit Precision

Kodai Ueyoshi, Masayuki Ikebe, Tetsuya Asai, Shinya Takamaeda-Yamazaki and Masato Motomura
Hokkaido University,
Graduate School of Information Science and Technology,
Sapporo, Hokkaido, 060-0814, Japan
Email: ueyoshi@lalsie.ist.hokudai.ac.jp, {ikebe, asai, takamaeda, motomura}@ist.hokudai.ac.jp

***Abstract*—Deep learning, especially the convolutional neural network (CNN), is a state-of-the-art model that can achieve significantly high accuracy in many machine learning tasks. Recently, efficient hardware platforms for accelerating CNN have been throughly studied. A binarized neural network has been reported to minimize the multipliers, which consume a large amount of resources, with a minimal decrease in accuracy. In this study, we analyzed the optimal performance of CNN implemented on an field programmable gate array (FPGA) considering its logic resources and a memory bandwidth, using multiple types of parallelisms such as kernels, pixels, and channels both in conventional and binarized CNNs. As a result, it became clear that all the parallelisms are required for the binarized neural network to obtain the best performance.**

I. Introduction

Convolutional neural network (CNN) has recently been gaining attention because of its impressive performance in various applications [1]. However, the performance requirements of practical applications cannot be achieved with a conventional CPU owing to the computational complexity of CNN. Therefore, hardware accelerators have been applied to obtain effective performace. Especially, FPGA accelerators are attracting more and more attention in this research field because of their good performance, high energy efficiency, development cost, and reconfigurability.

In previous studies, Zhang *et al*. focused on a structure that varied channels in each layer of CNN [2]. The study proposed the optimization method of implementation of CNN on an FPGA. On the other hand, several groups have studied facilitating a hardware implementation, such as a binarized neural network, which is able to reduce the computational resource used. Courbariaux *et al*. showed that replacing the multipliers with XNOR logics is possible by approximating the weights and inputs to binary with a minimal drop in accuracy [3], [4].

In this study, we analyzed the optimized accelerator for CNN in an environment that can be implemented with an increased number of neurons in the binarized neural network. As a result, we proved all parallelisms, which include channels, kernels, and output pixels not shown in previous works [2], [5], [6], can improve the performance of 8.38 Tera operation per second (TOPS) on a Virtex-7 FPGA when the number of implementable neurons increase by binarized CNNs.

II. Convolutional neural networks

A. Basics of convolutional neural networks

CNN has been gaining great attention in the field of computer vision because it is inspired by visual cortex. CNN is a multi-layered neural network with an input layer, multiple hidden layers, and an output layer. Within CNN, the hidden layers are referred to as convolutional layers that extract the input feature as feature maps. A subsampling layer is also included in the CNN hidden layers. However, a previous study has shown that convolutional operations occupy 90% of the computation time in the feed-forward process [7]. Therefore, we focused on accelerating the convolutional operations.

CNN employs a feed-forward process as an inference, and a backward process as training, similar to a classical multi-layer neural network. In practical application, the training is processed offline using high performance computing, and the inference is processed online to realize real-time processing. In this study, we treated the feed-forward process for online processing.

In this study, we considered AlexNet [8] as a benchmark to compare with previous studies [2], [5], [6].

III. Optimizing analysis

A. Optimization model

We accelerated the CNN operation to optimize the parallelism by layer, focusing on input and output channels, kernels, and output pixels. Figure 1 illustrates each parallelism. We defined the parallel parameters as follows: T_n as the input channel, T_m as the output channel, T_c as the output pixel of the column, T_r as the output pixel of the row, and T_k as the kernel. The parallel patterns we analyzed are as follows:

1) Channel parallelism (T_n, T_m)
2) Pixel parallelism (T_r, T_c, T_k)
3) Channel and pixel parallelism (T_n, T_m, T_r, T_c, T_k).

Figure 2 shows an overall view of our design. Each computing engine (CE) has a product-sum tree. The size of the CE (x) that is the number of multipliers, and the number of CEs (y) are parameterized by the parallel parameters shown in Table I.

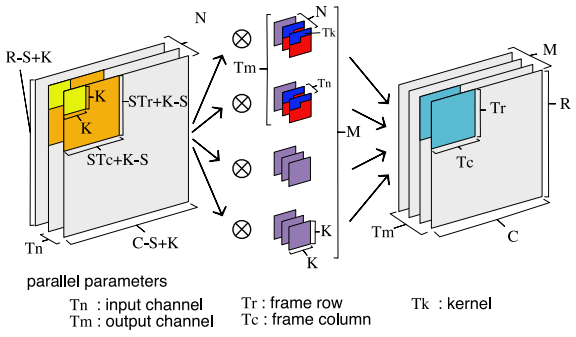


Fig. 1. Design model.

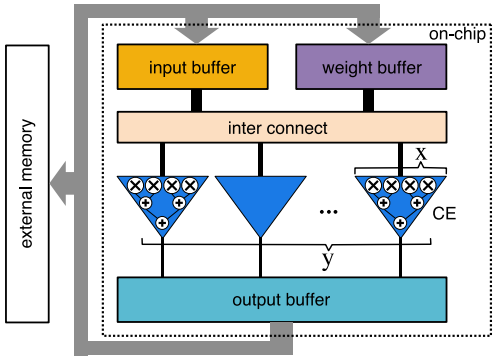


Fig. 2. Overall design.

TABLE I
THE SCALE PARAMETER OF SUM-PRODUCT UNITS.

Parallelism	x	y
Channel	T_n	T_m
Pixel	$T_k \times T_k$	$T_r \times T_c$
Channel+pixel	$T_n \times T_k \times T_k$	$T_m \times T_r \times T_c$

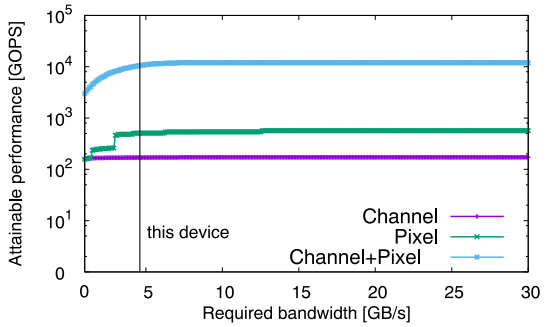


Fig. 3. Throughput of each parallelism depending on bandwidth using binarized CNN.

B. Analysis in the case of binarized CNN

In binarized CNN, the multipliers are not required in the convolutional layers. Therefore, the internal resource is determined by the number of slices, or logic elements, instead of the DSP units. In binarized CNN, the multipliers in CE can be replaced by XNOR logics [3], [4]. In these studies, a batch normalization [9] is used to maintain high accuracy. However, the calculation of the batch normalization can be omitted in the case of an inference [10]. Therefore, the adder trees occupy the most of resources. We analyzed the performance of binarized CNN using 80 % slices of maximum usable slices of a Virtex-7 FPGA implemented with Vivado HLS (v2015.2).

From the result shown in Fig. 3, the performance of the channel and pixel parallelism improved by 10.03 TOPS against the pixel parallelism, which achieved the second highest performance. From these results, the higher parallelism enhanced the performance significantly in the case of binarized CNN.

IV. Conclusion

In this paper, we analyzed the optimal parallel CNN accelerator using various parallelisms to obtain the best performance. As a result, we found that the channel and pixel parallelism could achieve the highest performance with low local memory capacity using DSP units in a Virtex-7 FPGA. In addition, we adapted it for binarized CNN that can replace the DSP with XNOR logic. As a result, we obtained the most efficient performance, which is 8.38 TOPS, by using binarized CNN.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[2] C. Zhang, P. Li, G. Sun, Y. Guan, B. Xiao, and J. Cong, "Optimizing fpga-based accelerator design for deep convolutional neural networks," in *Proceedings of the 2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, ser. FPGA '15. New York, NY, USA: ACM, 2015, pp. 161–170.

[3] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or -1," *arXiv:1602.02830*, 2016. [Online]. Available: http://arxiv.org/abs/1602.02830

[4] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "Xnor-net: Imagenet classification using binary convolutional neural networks," *arXiv:1603.05279*, 2016. [Online]. Available: http://arxiv.org/abs/1603.05279

[5] M. Motamedi, P. Gysel, V. Akella, and S. Ghiasi, "Design space exploration of fpga-based deep convolutional neural networks," in *2016 21st Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2016, pp. 575–580.

[6] A. Rahman, J. Lee, and K. Choi, "Efficient fpga acceleration of convolutional neural networks using logical-3d compute array," in *2016 Design, Automation Test in Europe Conference Exhibition (DATE)*, 2016, pp. 1393–1398.

[7] J. Cong and B. Xiao, *Minimizing Computation in Convolutional Neural Networks*. Cham: Springer International Publishing, 2014, pp. 281–290.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[9] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift." in *ICML*, ser. JMLR Workshop and Conference Proceedings, F. R. Bach and D. M. Blei, Eds., vol. 37. JMLR.org, 2015, pp. 448–456.

[10] H. I. Hiroki Nakahara, Haruyoshi Yonekawa and M. Motomura, "A batch normalization free binarized convolutional deep neural network on an fpga," in *25th ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (ISFPGA)*, 2017 (to appear).

Learning Interior Design using Convolutional Neural Networks

Xiaoxiong XING*, Yoshinori DOBASHI*, Tsuyoshi YAMAMOTO*
*Laboratory of Information Media Environment
Hokkaido University

***Abstract*—Previous works on interior design have used optimization applied to hand-crafted cost functions. There are works which design their cost functions by following interior design guidelines or through experience, and there are works that start by building statistical models reflecting furniture to furniture's spatial relationships and then sample from those models. Neural networks, on the other hand, excels at finding the intrinsic relationship among furniture in a design sample, therefore, we propose to apply convolutional neural networks to learning end-to-end interior design.**

I. Introduction

When people buy a new home, the first thing they would do before moving into it is to choose various furniture to furnish their home, which we call the process interior design. During years of practice, interior designers had gained experience, and some of them had summarized their experience as design guidelines. Merrell et al. [1] expressed the design guidelines as mathematical cost functions of furniture's positions and rotations, then the furniture's optimal positions and rotations were determined by minimizing those cost functions. Yu et al. [2] proposed a method similar to [1] by designing the cost functions themselves. Fisher et al. [3] proposed to model the relative position and rotation between two categories of furniture as Gaussian Mixture Models. They then fit those models using real design samples. At inference time, they first sampled from those models to initialize furniture's positions and rotations and then repeatedly adjust them to maximize the likelihood estimated from training samples.

Now we come to the era of Big Data. Our partner MyHome3D Corp., who released an online interior design application that users can use it to design their homes with tens of thousands of 3D furniture models provided by real world furniture companies. MyHome3D also released over one thousand design samples to assist users' design work and provided us with access to those data. Thanks to their design samples, we are able to apply deep neural networks to learning interior designs.

II. Method

When interior designers are asked to design a home, they will first get a floor plan then place furniture into it. Simulating this process, we propose to let the convolutional neural networks(CNN) to act like a designer who accepts a floor plan as its input and outputs a complete design.

Our method is inspired by [4] which used a CNN to produce an output image based on an input image. As their method follows the structure of generative adversarial networks(GAN), it is also called conditional adversarial networks. We refer the reader to their paper for a brief introduction to generative adversarial networks as it is out of the scope of this short paper.

Fig. 1. Top-down view of a design sample. The living room is highlighted with red outline.

A. Data Preparation

Fig. 1 is a top-down view of a design sample. In our research, we only consider the living room (most often join with the dinning room) as it contains the largest variation both in shape and furniture categories. In addition, we selectively consider the following nine furniture categories that occurs most often in design samples: sofa set, single-person sofa, two-person sofa, TV stand, tea table, cabinet, side table, dinning table set and chair. We extract from each design sample the coordinates of every wall's corners, the coordinate of every furniture's center and the rotation of every furniture.

When preparing the training sample, we scale all design samples with identical ratio to fit into a 16×16 grid. The length of a sample's last dimension is 19, of which 11 numbers represent the one-hot encoding of the above nine selected furniture categories plus wall and empty space, and the rest 8 numbers represent the one-hot encoding of the rotation angle from 0° to 360° as multiples of 45°. The bottom figure in Fig. 3 is an example of the training sample.

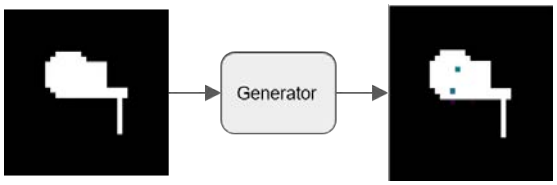


Fig. 2. Generator.

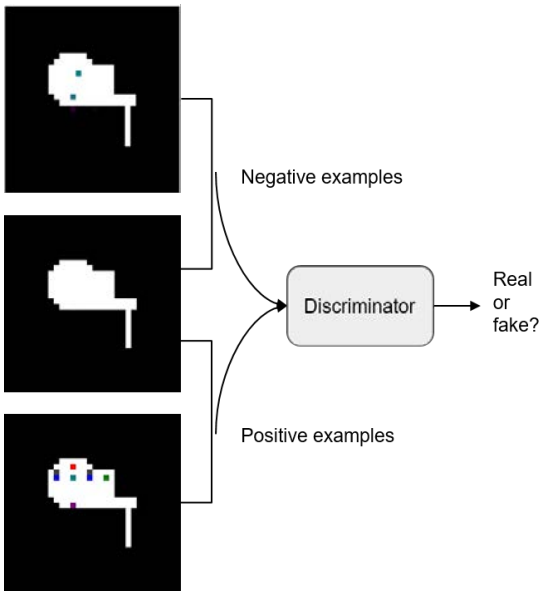


Fig. 3. Discriminator.

B. Neural Network Structure

We call the CNN that acts like the interior designer as Generator following the convention of GAN. It takes the floor plan as input and outputs a complete design. Following GAN, another CNN called Discriminator is used to learn to distinguish between real design samples and design samples generated by the Generator.

We use U-Net as the structure of the Generator and a three-layer CNN as the Discriminator. The Discriminator serves as a source of loss for the Generator. Besides that, as we only allow the Generator to put furniture within the room, we designed another loss (Eqn. 1) to prevent the shape of rooms from changing.

$$\begin{aligned}\text{cross\_entropy}(&\text{generated\_design} * (1 - \text{room\_mask}) + \\ &\text{room} * \text{room\_mask}, \text{room})\end{aligned} \tag{1}$$

During each iteration of training, a floor plan is fed into the Generator, and the Generator outputs a design (Fig. 2). The output of the Generator together with the floor plan are then fed into the Discriminator as negative examples. A training sample together with the floor plan are fed into the Discriminator as positive examples (Fig. 3).

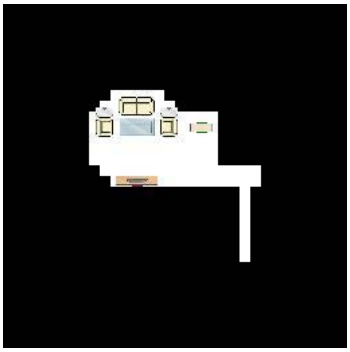


Fig. 4. An example of the real design sample used in evaluation.

III. Experiment and Evaluation

This research is still on going. We got 600 living room design samples with unique floor plan. 540 of them are used in training, and 60 of them are used in validation. Because it is difficult to evaluate an artistic design numerically so we plan to invite humans to do the evaluation in the future. We plan to present evaluators with an image of real design sample and an image of generated sample side by side, and ask them to choose one from the following there choices (1) The left image is a real design sample. (2) The right image is a real design sample. (3) Cannot distinguish. Our goal is to exceed 25% of the time that a user either chooses (2) or (3).

Every training sample and the predicted output of our neural networks are $32\times32\times19$ dimensional tensors. Simply mapping categories to colors as in Fig. 3 is not intuitive for evaluation, so we decided to draw a furniture at each furniture predicted location (Fig. 4).

Acknowledgment

We would like to thank MyHome3D Crop. (http://www.fuwo.com) for providing us with interior design samples. Xiaoxiong XING is supported by Global Station for Big Data and Cybersecuity (GSB) at Hokkaido University.

References

[1] P. Merrell, E. Schkufza, Z. Li, M. Agrawala, and V. Koltun, "Interactive furniture layout using interior design guidelines," in *ACM SIGGRAPH 2011 Papers*, ser. SIGGRAPH '11. New York, NY, USA: ACM, 2011, pp. 87:1–87:10. [Online]. Available: http://doi.acm.org/10.1145/1964921.1964982

[2] L.-F. Yu, S.-K. Yeung, C.-K. Tang, D. Terzopoulos, T. F. Chan, and S. J. Osher, "Make it home: Automatic optimization of furniture arrangement," in *ACM SIGGRAPH 2011 Papers*, ser. SIGGRAPH '11. New York, NY, USA: ACM, 2011, pp. 86:1–86:12. [Online]. Available: http://doi.acm.org/10.1145/1964921.1964981

[3] M. Fisher, D. Ritchie, M. Savva, T. Funkhouser, and P. Hanrahan, "Example-based synthesis of 3d object arrangements," *ACM Trans. Graph.*, vol. 31, no. 6, pp. 135:1–135:11, Nov. 2012. [Online]. Available: http://doi.acm.org/10.1145/2366145.2366154

[4] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *CoRR*, vol. abs/1611.07004, 2016. [Online]. Available: http://arxiv.org/abs/1611.07004

A Time-Division Multiplexing Ising Machine on FPGAs

Kasho Yamamoto, Shinya Takamaeda-Yamazaki, Masayuki Ikebe, Tetsuya Asai and Masato Motomura
Graduate School of IST, Hokkaido University
Kita 14, Nishi 9, Kita-ku, Sapporo 060-0814, Japan
(yamamoto@lalsie.ist.hokudai.ac.jp)

***Abstract*—machines based on the Ising model which can solve combinatorial optimization problems is an emerging solution to overcome the performance limit of von Neumann architecture. However, it is difficult to solve practical combinatorial optimization problems by existing approaches of FPGA-based annealing machines, due to the small number of implementable spins. In this paper, we propose the time-division multiplexing Ising machine architecture that efficiently utilizes on-chip memory resources in an FPGA, in order to address large scale combinatorial optimization problems. The evaluation result shows that it is possible to increase the spin number by 64 times compared to the conventional annealing machine.**

I. Introduction

Combinational optimization is a fundamental and practical method to describe and solve various social problems in our daily lives, such as transportation cost optimization. The difficult point to solve such optimization problems is that they are known as NP, so that they can be resolved in a polynomial time. Even if approximate solutions via heuristic approaches are allowed, it takes certainly long computing time to solve them due to their iterative searches of optimal solution points.

In order to overcome inefficiency of the modern von Neumann computers for such optimization problems, an Ising computer has been proposed based on the "natural computing" paradigm which maps a target problem onto a physical model in nature and observes the obtained status of the physical matters as its computing result. Especially, Yamaoka et al. has proposed a CMOS annealing LSI [2] to accelerate combinational optimization problems by utilizing artificial Ising model as electric circuits. Additionally, FPGA-based Ising machines are attractive alternatives for easy development of the systems instead of custom LSIs. One of the main problems of existing FPGA-based Ising machines is the limitation of simulated spin count coming from available hardware resource amount.

In this paper, we explore a novel FPGA-based Isign machine architecture for increasing the simulated spin count. We found that the prior FPGA-based implementations did not utilize on-chip memory blocks as known as "Block RAM" (or "BRAM" in short) in Xilinx FPGAs) on an FPGA effectively. Thus we focus on employing the on-chip memory block to increase the spin count. The contributions of this papers are described as follows:

1) We present a time-division multiplexing architecture of Ising machine on an FPGA that utilizes on-chip memory blocks for the spin count increase. In the prior works, a target Ising structure directly mapped on dedicated hardware spins; each spin is emulated by its own dedicated hardware spin unit. Therefore, the implemented spin count is directly limited by the size of the logic circuit resources. In our approach, we introduce an abstract model to separate the target Ising structure and the hardware structure. Statuses of spins are calculated and updated by time-division spin units with stream spin information suppliers of on-chip memory blocks. It enables to flexibly expand the spin count independently of the available logic circuit resources of an FPGA, such LUTs and registers.
2) We evaluated the efficiency of our proposal by using commercial FPGA synthesis tools. The evaluation results show that our architecture can handle 64 times more spin than previous one.

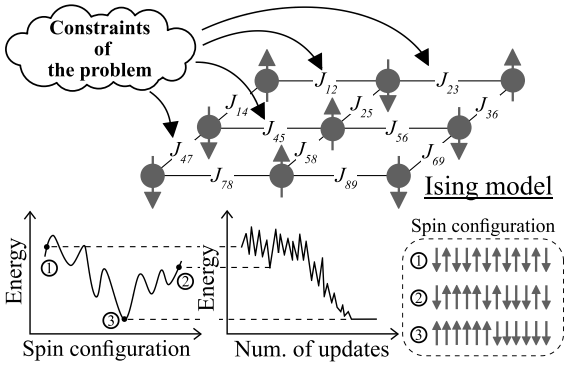


Fig. 1. Ising model

II. Ising Model

$$H = -\sum_{<i,j>} J_{ij}\sigma_i\sigma_j - \sum_i h_i\sigma_i \qquad (1)$$

The Ising model is a statistical model representing the behavior of the spins of a magnetic material. The energy function of the Ising model is represented by the equation (1), where σ_i are individual spin states, j_{ij} are the interaction coefficients that represent the strength of the interactions between different pairs of spin states, and h_i is the external magnetic coefficients. The Ising model has the property that the state of each spin is updated so that its energy function is minimized. Therefore, when a combinatorial optimization problem is mapped on to the Ising model, the combination of solution parameters that minimizes its energy can be observed naturally obtained after enough updates of spins as shown in figure 1.

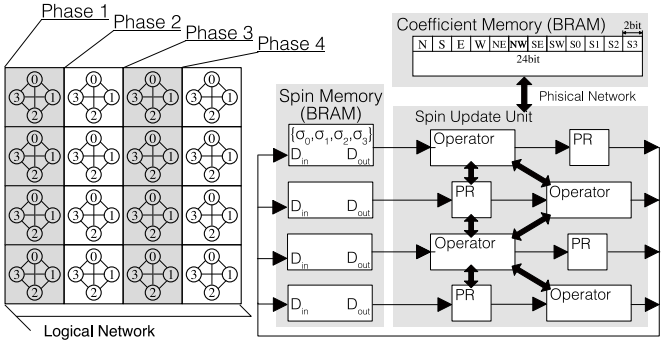


Fig. 2. Proposed Architecture

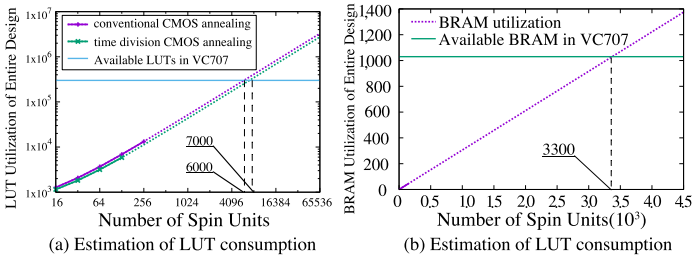


Fig. 3. Estimation of the number of the Spin Update Unit

III. Architecture

We propose a scalable architecture that processes large combinational optimization problems on limited hardware resources by separating spins of a target problem into both spacial and temporal directions. In this paper, we utilize the chimeric topology [1] for well known optimization problems.

Figure 2 presents the proposed architecture and processing mechanism. In the chimera topology, four complete spins are contained within one Spin Unit. Each internal spin is coupled to the same position spin inside the adjacent Spin Units.

The architecture consists of a spin memory (Spin Memory) that stores the state of all spins, an interaction coefficient memory (Coefficient Memory) that stores the interaction coefficient between adjacent spins, and Spin Update Unit which updates the state of spins row by row including Operator / Pipeline Register (PR) array, where the Operator determines the next state of the according to spin from the state of the adjacent spin and the interaction coefficient. In the Spin Update Unit , there is a constraint that adjacent spins cannot be updated simultaneously [2]. Therefore the pipeline registers are sandwiched between Operators.

The behavior of the architecture is as follows. As mentioned above, four complete spins (local spin) are contained within one spin unit in the chimera topology. This architecture updates sequentially from the local spin 0 in the chimera topology. A target Ising network is separated into 4 sections; We hereinafter call the section ”Phase”. First, spins in phase1 shown at the top of the figure 2 are read and transferred to the Spin Update Unit. the interaction coefficients of Phase 0 and Local spin 0 of Phase 2 are read from Spin Memory and Coefficient Memory, respectively. Then, they are transferred to the Spin Update Unit. Based on the read value, the Spin Update Unit updates the local spin 0 of phase 1. The back operator whose PR is inserted in the preceding stage receives the updated spin state from the front operator. This process is similarly performed up to Phase 4. After that, processing for local spin 1 is performed from phase 1 and it is repeated until all the local spins are updated. The architecture repeats this until the Ising model converges.

IV. Evaluation

Based on results of syntheses, we estimated the number of Spin Units that can be deployed on two architectures on the target board (FPGA : Virtex-7 XC7VX485T). The results for the LUT are shown in figure 3 (a), and figure 3 (b) shows the results for the BRAM. Note that the horizontal axis of figures is not the number of spins, but the Spin Unit (4 spins). The dotted line in the figure 3 is extrapolated based on a small scale implementation.

As can be seen from the figure, since the LUT consumption of the proposed architecture is smaller than that of the conventional method, the proposed architecture can allocate many Spin Units. However, even with the maximum use of 2000 BRAMs of 18 kb (1000 in the case of using 36 kb), there are about 3,300 Spin Units that can be deployed considering BRAM consumption in figure 3. Compared to the conventional architecture, our architecture has the same number of spins by dividing the process into two, and it can handle the spin number over the conventional architecture with division of three or more. Further, if the Ising network is divided by 128, our architecture can hold about 211,000 spins 64 times larger than the conventional architecture.

V. Conclusion

In this study, we proposed an annealing machine of the Ising model using BRAM dedicated to combinatorial optimization problem. Compared with the architecture of the previous research, the number of spins that can be processed at the same time is halved because the read bit width of the BRAM is limited. However, by maximizing the use of BRAM, it is possible to deploy 64 times the spin units of the previous research.

In the future, we will study hardware topology and partitioning method required for actual applications. We would like to improve our proposed architecture for more flexibility to the problem.

References

[1] C. Yoshimura, M. Hayashi, T. Okuyama, and M. Yamaoka, ”FPGA-based Annealing Processor for Ising Model”, International Symposium on Computing and Networking (CANDAR), 2016.

[2] M. Yamaoka, C. Yoshimura, M. Hayashi, T. Okuyama, H. Aoki, and H. Mizuno, “20k-spin Ising chip for combinational optimization problem with CMOS annealing,” ISSCC Dig. Tech. Papers, pp. 432 - 433, 2015.

Method of detecting errors and classifying mixed genomes based on the characteristics of reads orientation

Keito Aoki
Graduate School of Information Science and Technology
Hokkaido University
Sapporo, Japan
k-aoki@ist.hokudai.ac.jp

Kanako O. Koyanagi
Graduate School of Information Science and Technology
Hokkaido University
Sapporo, Japan
kkoyanag@ist.hokudai.ac.jp

Hidemi Watanabe
Graduate School of Information Science and Technology
Hokkaido University
Sapporo, Japan
watanabe@ist.hokudai.ac.jp

***Abstract*—Next generation sequencing (NGS) achieved producing several billion short DNA sequences or reads. However, the high average error rate of such reads makes it difficult to find rare target sequence changes, e.g., those causing cancer development and different phenotypes between races. The purpose of this study is to detect and correct the sequencing errors as well as to separate mixed genomes. We develop an algorithm meeting this purpose by looking at the base frequencies between differently oriented reads from the same DNA position. We tested the algorithm using data sets of human adenovirus genomic data that we have sequenced using a next generation sequencer.**

Keywords—DNA sequence, Next generation sequencing, Error detection, statistical analysis

I. INTRODUCTION

Next generation sequencing (NGS) has enabled us to sequence DNA bases faster and cheaper than traditional Sanger method. This advantage is achieved by producing several billion short DNA sequences (reads) in parallel [1]. To analyze and to get more information from this big data, we need apply efficient and effective algorithms.

One of the disadvantages of NGS is the higher error rate (10^{-2} - $4*10^{-2}$) than that of Sanger method (10^{-2} - 10^{-2}) [2]. For this reason, it is difficult to distinguish errors from substitutions or polymorphic sites. In general, majority vote by higher coverage, assumption of probability distribution of errors, specialized library preparation are used to remove errors [3,4,5]. However, if samples contain heterogeneous genomes such as genomes of closely related species, we would exclude a genome with low frequency because these sequences are determined as error.

The goal of this study is to develop a method, which can distinguish errors from substitutions or polymorphic sites. We first determine errors in each site by testing asymmetric distribution of reads (A), and then classify reads into the original genomes they belong (B).

II. METHODS

We applied this algorithm to a real data set (Human Adenovirus (HAdV) species B), which was sequenced by Ion Torrent Personal Genome Machine (PGM) [6]. After 3' quality clipping was applied to reads by FASTX-Toolkit [7], reads were mapped to reference genomes by Bowtie2 [8]

A. Error Detection

We used not only the frequency of bases, but also the orientation of reads for detecting errors. DNA were fragmented and sequenced randomly in fragments and direction. Thus, reads should be distributed randomly across a genome in the forward and reverse orientation. However, errors would exist in asymmetric orientation in which errors exist more in forward reads than in reverse reads or vice versa. And errors were not changed into same other bases among reads and reads orientation.

Thus, we generated 2 * 4 contingency table, forward/reverse versus Adenine/Thymine/Guanine/Cytosine for each site (**TABLE I**). And then, we performed Chi-squared test for every site, and tested the null hypothesis that the frequency of bases at the site is not asymmetry. If the null hypothesis was rejected, we determined the site included errors.

TABLE I. Contingency Table of Site i

Read Orientation	**The Frequency of Base**			
	Adenine	Thymine	Guanine	Cytosine
Forward	$F_A(i)$	$F_T(i)$	$F_G(i)$	$F_C(i)$
Reverse	$R_T(i)$	$R_A(i)$	$R_C(i)$	$R_G(i)$

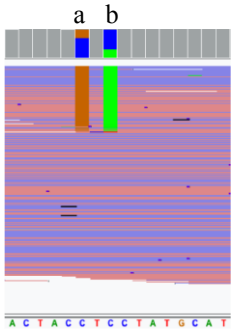


Fig. I: This shows NGS reads on IGV (Integrative Genomics Viewer). The red horizontal bars are forward reads and blue bar are reverse reads. The places changing color have different bases with reference genome. The vertical width indicate coverage of each sites.

B. Reads Classification

After removing errors by above procedure (A), we cluster sites that included the same base frequencies of segregating for classifying reads. For example, site "a" and site "b" had approximately same base frequencies of segregating sites, so that we could classify reads in which two sites from same frequencies were included (**Fig. I**). To indicate such relation, we clustered sites by generating 2 * 4 contingency table, order of frequency ($f_1(i)$, $f_2(i)$, $f_3(i)$, $f_4(i)$), and then used Chi-squared test for all the combinations of those sites. Finally, we classified reads in which two or more sites from same classification were included.

III. RESULTS

In previous process, we used HAdV specied B as a reference genome and we obtained mapping read sufficiently (**TABLE II**).

As a result of method A, 884/35239 sites were detected as error including sites. And then, we applied method B to remaining 34355/35239 sites, and classified into two groups based on the base frequencies of segregating sites. One of the group contained 109/34355 sites which were showed average ratio of $f_1(i)/f_2(i) = 4.18$ suggesting.

Consequently, 4663 reads were classified into two different genomes by method B.

IV. DISCUSSION

In the result of method A, there were some characteristics that detected sites were often nearby homopolymer and had asymmetrical indels (insertion or deletion). This is known as characteristics due to the principle of Ion PGM [4]. Other than that, there were also significantly higher error or characteristics at some sites. For example, there were far from homopolymer and the quality score was very low. We suggested that there were some difference among genome region.

By using this algorithm, we could detect sites including errors and classify reads into original genomes. In the case of this sample, we suggested that the mixed genome were very close, so that segregating sites were not many. However, classified reads were not searched as closely related species. This indicates this sample didn't include more than one species, but the genome had some polymorphism.

In this study, we used a raw data of HAdV on Ion PGM. But this algorithm would be applied to other platforms and species.

TABLE II. Sample Data

Species	Reads	Reads Used	Reads mapping	Mean length	Mean quality
HAdV_B	34,027	32,458	95.40%	246.8	27.04

In this study, we mapped reads to the reference genome for assembling. However, this algorithm only needs the frequency of bases and does not need the information of position on the genome. Thus, we may remove the restriction of necessity of reference genome and influence of mapping tool's error for future work.

REFERENCES

[1] S. Goodwin, J. D. McPherson and W. R. McCombie, "Coming of age: ten years of next-generation sequencing technologies," *Nautre Review Genetics*, 2016, vol. 17, 333-351

[2] T. C. GLENN, "Field guide to next-generation DNA sequencers," *Molecular Ecology Resources*, 2011, vol. 11, 759-769

[3] X. Yang, S. P. Chockalingam and S. Aluru, "A survet of error-correction methods for next-generation sequencing," *Briefings in Bioinformatics*, 2013, vol 14, 56-66

[4] L. M. Bragg, G. Stone, M. K. Butler, P. Hugenholtz, G. W. Tyson, "Shining a Light on Dark Sequencing: Characterising Errors in Ion Torrent PGM Data," *Computional Biology*, 2013, vol. 9, Issue. 4

[5] D. H. Spencer, M. Tyagi, F. Vallania, A. J. Bredemeyer, J. D. Pfeifer, R. D. Mitra, and E. J. Duncavage, "Performance of common analysis methods for detecting low-frequency single nucleotide variants in targeted next-generation sequence data," *The Journal of Molecular Diagnostics*, 2014, vol. 16, 75-88

[6] J. M. Rothberg, W Hinz, T. M. Rearick, 3rd ed, "An integrated semiconductor device enabling non-optical genome sequencing," *Nature*, vol. 475, 348-352

[7] Gordon, Assaf and G.Hannon, " Fastx-toolkit. FASTQ/A short-reads pre-processing tools," *Unpublished Available online at: http://hannonlab.cshi.edu/fastx_toolkit*, 2010

[8] B. Langmead and S. L. Salzbarg, "Fast gaped-read alignment with Bowtie 2," *Nature Methods*, 2012, vol. 9, 357-359

Amino Acid Exchangeability and Disease-causing Ability in Human Beta Globin Gene

Sangeetha Ratnayake*, Toshinori Endo*, Naoki Osada*

* Division of Bioengineering and Bioinformatics,
Graduate school of Information Science and Technology,
Hokkaido University,
Hokkaido, Japan

***Abstract*- The effect of amino acid changes on phenotypes, whether particular mutations cause disease or not, became important in therapeutic medical investigations. Many studies have been carried out considering the information about evolutionary conservation, stability of the protein, or physiochemical properties of amino acids to understand the relationship between mutations and their consequences. In order to understand the relationship, we focus on the human beta globin gene (HBB). HBB is an important subunit along with alpha subunit and composes hemoglobin protein, which plays a vital role in humans as it transports oxygen and other gases throughout a body. Disorders in HBB are one of the most frequently observed genetic diseases in humans, where many mutations have been reported at almost every amino acid site. In this report, we try to elucidate the relationship between disease causality and properties of amino acid changes, in order to understand the reason behind the cause of hereditary genetic diseases like beta thalassemia. We found that physiochemical property of amino acid changes plays the most important role for predicting disease causality in HBB, yet complex relationship between physiochemical groupings does exist.**

Key words- Evolutionary conservation, Physiochemical properties, HBB

I. INTRODUCTION

Globin protein family, is a large and well-studied family that has widely distributed in many organisms. Vertebrates have different types of hemoglobin in different stages of life. Disorders of the Hemoglobin are one of the most common inherited diseases in humans. Many mutations in almost every amino acid position in α-globin and β-globin, which effect for the structural and functional behaviors of hemoglobin in humans have been reported. For instance, sickle cell disease and thalassemia can be found as genetic diseases in some populations. A database to study about the hemoglobin variants is Globin Gene Server (http://globin.cse.psu.edu/). Total or partial absence of β-globin causes β thalassemia, the most common thalassemia variant in a human population, where the malaria was epidemic [1].

Evolutionary conservation is a key to understand disease causality of amino acid changes in genomes. However, in some cases, evolutionary information is not sufficient to understand the disease causality. In such situations, structural and functional stability of the protein structure, physiochemical properties such as polarity, hydrophobicity and the size of the amino acids can be accountable to a better prediction.

II. MATERIALS AND METHODS

A. Protein structure

Three variables, protein stability, residue depth, and distance to iron molecule were measured using a protein structure in PDB database (PDB accession: 2DN1).

Although proteins with a sufficient stability in their native state function well and will have an identical fitness, a mutation could destabilize the both functional and structural behavior of the protein [3]. We used FoldX software to predict the change of stability [6]. Residue depth represents solvent exposure of the amino acid residues and indicates whether a specific site is buried in the space of protein structure or exposed to the surface [6]. As the hemoglobin is all about metal binding and gas transportation throughout the human body, heme pocket is the important feature of each globin proteins. We measured the distance to iron molecule from each amino acid site [6].

B. Miyata's classification of amino acids

Amino acids grouping was used to understand the relationship of amino acid exchangeability towards disease causality. Miyata's classification represents the distance between groups of amino acids. This measurement was used as a variable to predict the disease causality [2].

C. Evolutionary conservation

Protein sequences of vertebrate HBB were obtained from Ensembl, which has an automated pipeline for retrieve genes and annotation based on protein and mRNA evidence in UniPortKB and NCBI RefSeq databases, along with manual annotation from the VEGA/Havana group [4]. To evaluate the evolutionary conservation level, Shannon entropy value was calculated using a multiple sequence alignment, based on a reconstructed gene tree of 30 vertebrates including, mammals (20), aves (3), reptiles (2), amphibians (1), bony fishes (3) and chondrichthyes (1) [8].

D. Logistic regression analysis

We applied logistic regression analysis to predict disease causality of amino acid changes in HBB [7]. Five independent variables were selected considering the evolutionary conservation, structure stability, and the physiochemical properties.

III. RESULTS AND DISCUSSION

Table 1: *P*-values obtained in the logistic regression analysis using five independent variables.

	Single variable significance				
Phy: Prop	0.0237				
Conservation	0.7019	0.0270			
Dist: to iron	0.9172	0.2018	0.8341		
Stability	0.7933	0.2706	0.0303	0.0229	
Resid: Depth	0.9373	0.2039	0.0900	0.5062	**0.0004**
		Phy: Prop	Conserv -ation	Dist: to iron	Stability
		Interacting variable significance			

*Physiochemical property (Miyata's classification), Evolutionary conservation, Molecular distance to iron molecule, structural stability and the solvent exposure. Red colored cells represent the significant variables.

Through searching databases and literatures, we identified 1685 mutations in human HBB. Among them, 1135 were synonymous mutations and 525 were non-synonymous mutations. Out of 525 non-synonymous mutations, 230 are known to cause disease such as beta thalassemia and sickle cell anemia.

The goal of logistic regression is to find the best fitting (biologically reasonable) model to infer the relationship between binomial characteristics on disease causality and the independent variables [7]. Except for the physiochemical property, none of the parameters showed any significance as a single variable. However, several variables showed significant effect when combined with other variables (Table 1): residue depth and molecular distance to an ion molecule. Interaction between stability and the residue depth showed smallest *p*-value ($p = 0.0004$).

Understanding of the pattern of disease causality heterogeneity based on all above variables is much complicated as there are many disease varieties in beta globin gene. Therefore, we categorized the disease causality using Miyata's classification of amino acids and the consistency of the pattern were tested using the cross validation. In addition to the result of logistic regression, we identified specific pattern in the physiochemical grouping of mutations, where specific type of amino acid mutations always cause disease, suggesting that there are some hidden reasons determining the disease causality of amino acid mutations (Figure 1). In terms of physiochemical differences, doubt has arisen with that certain mutations always caused non-disease variants even their differences are significant.

IV. CONCLUSION

Simple concept of physiochemical properties of amino acids effectively predict disease causality of HBB than evolutionary conservation. Logistic regression analysis suggested along with physiochemical properties structural stability are influential to disease causality only combined with the solvent exposure.

Amino acid exchangeability		Mutation amino acid					
		G1	G2	G3A	G3B	G4A	G4B
Wild type amino acid	G1	No data	No data	No data	100%	No data	100%
	G2	100%	100%	100%	100%	100%	60%
	G3A	No data	100%	65%	100%	75%	100%
	G3B	No data	95%	100%	65%	75%	100%
	G4A	No data	100%	95%	100%	100%	100%
	G4B	100%	100%	65%	100%	75%	100%

Figure 1: Amino acid exchangeability pattern of the dataset. Colors represent the disease (yellow) and non-disease (blue). The number represents the success rate of the prediction.

Amino acid exchangeability of disease causing mutations based on physiochemical groupings have shown a complex but consistent pattern, including exceptional amino acid changes of non-disease causing mutations. This realized pattern requires comprehensive realization of observed exceptional amino acid changes to improve its accuracy. In future, these consequences could be applied to improve the modern medical investigations such as predict disease causality of new mutations or therapeutic designing.

V. REFERENCES

[1]. Wild B and Bain B.J, Investigation of abnormal haemoglobins and thalassemia. Dacie and Lewis Practicle Haematology, Elsevier Health Sciences, UK, p. 271-309 2009.

[2]. Miyata T, Miyazawa S and Yasunga T,Two types of amino acid substitutions in protein evolution, J Mol Evol,12(3):219-36, 1979.

[3]. Echave J, Speielman S.J and Wilke C.O, Causes of evolutionary rate variation among protein sites. Macmillan Publishers Limited, Nature (Reviews), *Genetics,* p.109-121, 2016.

[4]. Cunningham F, Amode M. R, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, *et al*. **Ensembl 2015** *Nucleic Acids Research* 201543 Database, issue:D662-D669, doi:10.1093/nar/gku1010, http://asia.ensembl.org/index.html?redirect=no

[5]. Kuan Pern Tan, Thanh Binh Nguyen, Siddharth Patel, Raghavan Varadarajan and M. S. Madhusudhan, Depth: a web server to compute depth, cavity sizes, detect potential small-molecule ligand-binding cavities and predict the pKa of ionizable residues in proteins, Nucl. Acids Res. 41 (W1): W314-W321. doi: 10.1093/nar/gkt503, 2013.

[6]. Van Durme, Joost, Delgado Javier, Stricher Francois, Serrano Luis, Schymkowitz Joost, and Rousseau Frederic (2011), A graphical interface for the FoldX forcefield., Bioinformatics, Volume 27, Issue 12, p.1711-2

[7]. Frank Harrell (2015), Regression modeling strategies: with applications to Linear models logistic and ordinal regression, and survival analysis, Second edition Springer: ISBN 978-3-19425-7(eBook), p.219-307

[8]. R. R. Coifman and M. V. Wickerhausez, Entropy-based algorithms for best basis selection, in *IEEE Transactions on Information Theory*, vol. 38, no.2, pp.713-718, doi: 10.1109/18.119732, 1992.

American Traditional Bottle Gourds Possessed Hybrid DNA in the Nucleus and Chloroplasts: Alternative Scenario for Ancient Propagation of *Lagenaria siceraria*

Dai Watabe, Naoki Osada, and Toshinori Endo
Graduate School of Information Science and Technology,
Hokkaido University
Sapporo, Japan

Hiroshi Yuasa
The Research Institute of Evolutionary Biology
Tokyo, Japan

Keywords— bottle gourd, ancient cultivation, and DNA

I. INTRODUCTION

Bottle gourd *Lagenaria siceraria* is one of the oldest cultivated plant. In Mexico, ancient rind of the gourds were found at Guila Naquitz site dated back to 10k BP [1]. In Japan, ancient bottle gourd seeds were found at Lake Biwa site dated back to 9.6k BP [2]. On the other hand, the oldest African archaeological record dated back merely to 4,000 years in Egypt [3], although Africa is considered as the origin because all wild relatives are found exclusively in this continent [4, 5]. In regard to American bottle gourd origin, two hypothesis had been proposed. Erickson *et al.* (2005) concluded that ancient Asian gourds were brought to Americas by Paleoindians based on two chloroplast INDELs and one SNP [6]. Kistler *et al.* (2014) reported that ancient African bottle gourds had arrived in Americas based on whole chloroplast DNAs [7]. Results of Kistler *et al.,* however, had inconsistent INDELs pattern in terms of subtype classification result of Erickson *et al*. Therefore, we analyzed both chloroplast and nuclear DNAs to clarify the origin and propagation route of American bottle gourds.

II. METHODS

A. Seed samples

The following seed samples were provided by from Hiroshi Yuasa who collected them from local tribes at each site. In total there were thirty-nine Asian, twenty-four African, ten New Guinean, and nine American specimens. In addition, two Japanese samples (#J1 and #J2) were purchased from Takii Nursery Company and Ohta Nursery Company, respectively.

B. DNA extraction

Of the 70 seed samples, 46 germinated in ten days following rhizogenesis. DNA was extracted from the leaves after one month using the CTAB method [8]. For 14 samples, DNA was extracted directly from the seeds using a DNA Suisui-L Kit (Rizo Co., Ltd.).

C. DNA sequencing

DNA samples were sequenced by Hokkaido System Science Co., Ltd. using a BigDye Terminator v 3.1 Cycle Sequencing Kit, ABI PRISM 3130 Genetic Analyzer and ABI PRISM 3730 DNA Analyzer. The samples were sequenced using the dideoxy chain termination method.

D. Classification by variants

Because the probability insertions/deletions (INDELs) are lower than the probability of the SNPs, we assumed that the variations in these sequence are important. Therefore, in this study, we first evaluated INDELs and then evaluated SNPs.

III. RESULT

A. New primer design for PCR amplification

A total of 1,644 nucleotide sequences of the Cucurbitaceae species were downloaded from GenBank and clustered using BLASTClust to yield 28 alignments, which included at least five sequences. We selected variable DNA regions within the same species to design nine PCR primers.

Among nine PCR primers designed, we successfully amplified DNA fragments using seven primer sets *(trn*L-*trn*F, *rpl*20-*rps*12, *matK*, *rbcL*, *18S rRNA, 18SrRNA* upstream region and *GLY*). As a result, we used two primers (matK, *18SrRNA* upstream region) which variable among samples and five ready-made primers (*trn*C-*trn*D, *trn*S-*trn*G, BOP19_35, BR01_19 and LSR088) for PCR amplification.

Insertion/Deletions (INDELs) and Single Nucleotide Polymorphisms (SNPs) for sub-classification of African and Asian types

The INDELs in the *trn*C-*trn*D and *trn*S-*trn*G clearly classified samples into Asian and African subtypes, with the exception of #518 Ethiopia specimen being classified as Asian. For three chloroplast DNAs and four nuclear DNAs, SNPs also clearly classified samples into Asian and African types. Exceptionally, Ethiopian sample showed Asian types for BOP19_35, *matK* and *trn*C-t*rn*D. #371 New Guinea possessed the African-specific SNP and #366 New Guinea possessed the heterozygous African/Asian SNP for *18SrRNA* upstream region.

B. Nuclear sequences from the American samples contained African and Asian variants, except the Guatemalan gourd.

All of the American samples contained the African type of the INDELs in the chloroplast *trn*S-*trn*G and *trn*C-*trn*D region, except #277 Guatemala and #309 Mexico, which possessed the Asian type. All of the American samples also contained African type of chloroplast SNPs except #277 Guatemala and #309 Mexico, which possessed the Asian type. For nuclear SNPs, All of the American samples also showed hybrid type except #277 Guatemala which showed pure Asian type. For LSR088 and *18SrRNA* upstream region, all American samples showed Asian type. For BR01, all American samples showed African type except #277 Guatemala. For BOP19_35, five American samples showed Asian type and the remaining three American samples showed African type (Fig. 1).

IV. Discussion

The chloroplast INDELs and SNPs clearly separated the samples into Asian and African types. These results are consistent with the result of Erickson *et al.* [6].

The Ethiopian samples possessed Asian types of the chloroplast INDELs. These results are consistent with phylogenetic tree of Kistler *et al.* [7] and suggest that the East African gourds from Ethiopia propagated to Asia.

Chloroplast INDELs separated American samples into Asian and African subtypes except Mexico and Guatemalan samples. In nuclear DNA analysis, all American specimens were hybrids except one Guatemalan was pure Asian type, no pure African subtype was found. The African variants found in the American gourds suggest that the American gourds propagated from Africa. It can be speculated that American bottle gourds were brought from Africa by slave ships since the 16th century rather than by ocean current, since the rind of the wild ancestor of bottle gourds appear too fragile to cross

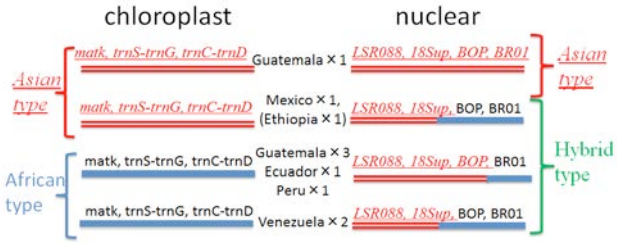


Fig.1 SNPs type of American and Ethiopian samples

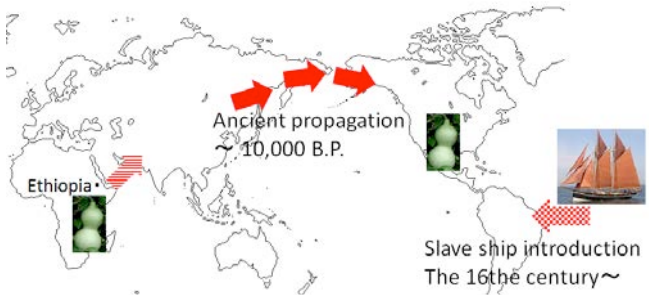


Fig. 2 Estimated bottle gourd propagation route

the Atlantic Ocean. Furthermore our samples were derived from the in-tribe grown and that no wild species were found in America. Therefore Asian variants found in the American gourds would suggest the ancient transmission to America predates by human carriage from Asia, rather than direct floating from Africa.

V. Conclusion

1. The chloroplast INDELs and SNPs clearly separated the samples into Asian and African types.
2. INDELs data suggested that the East African gourds from Ethiopia propagated to Asia. Our results are consistent with the result of Erickson *et al.* [6].
3. Hybrid American bottle gourds and pure Asian type Guatemalan gourds suggests American gourds were derived from Asia by human carriage.

References

[1] Smith BD, "*Cultural Evolution: Contemporary Viewpoints*", eds Feinman GM, Manzanilla M (Kluwer / Plenum, New York), 2000, pp. 15–60.

[2] Tsuji S, Nakamura T, Minaki M, Ueda Y, Kosugi M (1992) Channel of South Lake Awazu (2): Excavation survey summary report due to dredging. Awazu submarine archaeological site. eds Shigaken board of education, Shigaken Cultual propaties protection associations, pp 56–61.
辻誠一郎、中村俊夫、南木睦彦、小杉正人"南湖栗津航路(2)浚渫工事に伴う発掘調査概要報告書　栗津湖底遺跡", 滋賀県教育委員会・（財）滋賀県文化財保護協会編, 1992 年

[3] Schweinfurth G, "Futher discoveries in the flora of ancient Egypt", *Nature* 29:312–315, 1884

[4] Heiser CB, "The Gourd Book" University of Oklahoma Press, Norman , 1979, pp. 71–87.

[5] Heiser CB, "Foraging and Farming: The Evolution of Plant Exploitation", eds Harris DR, Hillman GC, Unwin Hyman, London, 1989, pp. 471–480.

[6] Erickson DL, Smith BD, Clarke AC, Sandweiss DH, Tuross N, "An Asian origin for a 10,000-year-old domesticated plant in the Americas", *Proc Natl Acad Sci USA* 102(51), pp. 18315–18320, 2005

[7] Kistler L, *et al.*, "Transoceanic drift and the domestication of African bottle gourds in the Americas" *Proc Natl Acad Sci USA* 111(8), pp. 1–5, 2014

[8] Doyle JJ, Doyle JL, "Isolation of plant DNA from fresh tissue." Focus (Madison) 12, pp. 13–15, 1990.