

Prediction of Compression Ratio for Transform-based Lossy Compression in Time-series Datasets

1st Aekyung Moon
ETRI, Korea, akmoon@etri.re.kr

2nd Juyoung Park
ETRI, Korea, jypark@etri.re.kr

3rd Yun Jeong Song
ETRI, Korea, yjsong@etri.re.kr

Abstract—As many IoT devices generate an enormous and varied amount of data that need to be processed in a very short time, storing and processing IoT big data become a huge challenge. While lossy compression can dramatically reduce data volume, finding an optimal balance between volume reduction and information loss is not an easy task. The compression ratio is within a range tolerable by the application is crucial. Motivated by this, we analyze the characteristics of data compressed and present a prediction model about the compression ratio of transformation-based lossy compression algorithms for IoT datasets collected.

Keywords:IoT(Internet of Thing), Data analytics, lossy compression

I. INTRODUCTION

In developing IoT applications, efficient management of big data is one of the challenging problems. Transmitting a large volume of data is one of the major causes of high energy consumption and high communication bandwidth usage [1], [2]. Data compression techniques can mitigate these problems since the storage and communication overheads are reduced. There are two types of compression techniques, lossy and lossless. Lossy compression can generate the reconstructed datasets with the loss of some data to its original datasets, on the other hand, lossless compression can generate the reconstructed datasets without any loss of data to its original form. Various lossy compression algorithms have been proposed to meet these big data management needs in the edge side of IoT.

Our motivation in this paper is to present a prediction model for transform-based lossy compression techniques. Naively applying transform-based lossy compression seldom gives a reasonably high compression ratio because of the inherent randomness exhibited in many time-series datasets. Because of this, combining data transformation with compression can be more effective as the transformed data usually reveal the correlation of the data explicitly. Even though, these lossy compression [3] reduces the data size significantly, the error introduced by applying transform-based lossy compression methods [3]–[5]. Therefore, compression schemes might be useful if they are applied to the more compressible variables ensuring data reliability. Under those backgrounds, we apply transform-based lossy compressions to time-series datasets and approximate them according to energy compaction. To

this compression process, we propose the prediction model of compression ratio for a transform-based lossy algorithm. DCT-K uses DCT (discrete cosine transform) to transform original data to coefficients values for enabling approximation. We focus on the K -dominant coefficients according to the fixed energy. In our evaluation, we use a real-world dataset from IoT stations located in an orchard in Youngcheon, South Korea. Our evaluation uses the following five most important variables: air temperature (TP), grass temperature (GTP), relative humidity (R-HMT), wind speed (WS), and soil humidity (S-HMT) in all environmental data. The data were continuously monitored and collected every minute using measuring (or sampling) periods.

II. RELATES WORKS

In [6], they compared the existing four lossy compression techniques using the signal characteristics of time-series sensor data to identify which compression technique achieves a higher compression ratio with minimal error rates. Due to the considerable impact of signal transform techniques, numerous scientific studies have focused on addressing the transform-based lossy compression algorithms. For instance, [7], [8] compared DCT, discrete wavelet transform (DWT), and other transformation algorithms for their applicability on Electrocardiogram (ECG) dataset. In [9], data transform techniques are compared for weather data. In [3] proposed DCT-based lossy compression algorithms, which achieves comparable performance, showing 3x–38x compression ratio while guaranteeing user-specified error bounds on the evaluated datasets. However, if a compression ratio reaches high, the reconstructed data are often distorted too much to tolerate.

In [10], by adapting to the compression speed of a host CPU, current system load, and network speed, they propose algorithms to estimate the compression time of each data block given in case of lossless compression. On the other hand, [11] describes compression ratio depends upon the percentage of zeros for DCT coefficients after quantization for DCT-based lossy compression in image datasets.

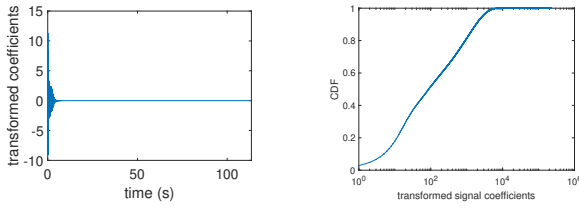


Fig. 1. The transformed representations of ECG signals. (a) after applying DCT transform. (b) Cumulative distribution function (CDF) in the sequences of transformed DCT coefficients.

III. PREDICTION COMPRESSION RATIO OF TRANSFORM-BASED LOSSY COMPRESSION

DCT, DWT, and Fast Walsh-Hadamard (FWHT) can be applied in a lossy compressor [2]. We select DCT (the most commonly used DCT-II) like [3], [12]. In transform-based lossy compression algorithms [13], unlike original signals, the transformed signals often contain *only* a few significant components [5], [14], [15]. In other words, discrete transforms redistribute the energy contained in the signal and concentrate it into a small number of dominant coefficients (as low-frequency coefficients).

To demonstrate that similar signal property of high energy compaction exhibits in ECG datasets, Figure 1 shows the distribution of coefficients (in the time domain) after applying a discrete transform (DCT in this case). Figure 1 depicts a scenario where DCT components containing 99%-percentile of the energy attained by the original data are only 2% (4,667 out of 232,400) of the entire ECG datasets. Transform-based lossy compression uses the energy spectrum of signal theory, where a given signal is the sum of the squared transform coefficients [5], [16]. While using the advantages of transformation mechanics, the objective of our paper is to present a prediction model of compression ratio for transform-based lossy compressor, namely, DCT-K.

A. Models

To describe our prediction model, we assume that the data point $X_t^{(i)} \in \mathbb{R}$, where $1 \leq t \leq n$ and $1 \leq i \leq p$, and t and i are the relative time in compressed periods (ultimately decomposed block in our case) and index of the compressed period, respectively. It also requires a parameter, $\theta^{(i)}$, which characterize data points in the compressed period. Each compressed period i has a series of data point during time t ($1 \leq t \leq n$). Let $\theta^{(i)}$ denote the parameter associated with typical characteristics of n series data (one period) in the period i . Each period i includes datasets $X_{(t)}^{(i)}$ follows the condition in Equation 1. The k -th block starts at s_k and ends at e_k .

$$X_{(t)}^{(i)} = \begin{cases} X_{(t)}^{(1)}, & \text{if } s_1 \leq t \leq e_1 \\ X_{(t)}^{(2)}, & \text{if } s_2 \leq t \leq e_2 \\ \dots & \\ X_{(t)}^{(p)}, & \text{if } s_p \leq t \leq e_p, \end{cases}, 1 \leq i \leq p. \quad (1)$$

We next define the data profile ($\theta^{(i)}$), which depends on whether the block under consideration is associated with data of typical or anomalous behavior. $\theta^{(i)}$ for characterizing data data points for each block i using Equation 2 below. Each block i has a series of features during time t ($1 \leq t \leq n$). Let $\theta^{(i)}$ denote the parameter associated with typical behavior of n series data (one block) to represent the characteristics of datasets in the block i .

$$\theta^{(i)} = \begin{cases} \theta_{(1)}^{(1)}, \dots, \theta_{(n)}^{(1)} \\ \theta_{(1)}^{(2)}, \dots, \theta_{(n)}^{(2)} \\ \dots \\ \theta_{(1)}^{(p)}, \dots, \theta_{(n)}^{(p)} \end{cases}, 1 \leq i \leq p. \quad (2)$$

To select feature for characterize datasets and related to the transformation compression, we choose the following metrics to assess the characterization of the datasets to be compressed.

- Variance (x), mean (x) and median (x), skewness (x) and kurtosis (x) of dataset can be calculated.
- Stationarity means that the statistical properties of a time series do not change over time. Therefore, the ability to determine whether a time series is stationary and making effective and precise predictions are important. We use two statistical tests would be used to check the stationarity of a time series – the Augmented Dickey-Fuller (ADF) test and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test.
- Calculate the tn -th discrete difference along the given period i . The first difference is given by $diff(t) = X_{t+1}^i - X_t^i$ along the i period, higher differences are calculated by using diff recursively. we calculate the standard deviation of (diff(t)).

Selecting an appropriate algorithm is essential when the training data is ready to use after the preprocessing, which involves balancing the class distribution of given data, normalization, and missing value analysis. However, all learning algorithms perform equally well when averaged over all possible data sets. In other words, it meant that looking for a general and superior algorithm to other algorithms is not feasible. Therefore, we use machine learning algorithms to predict compression ratio. There are six representative regression methods: DT (decision tree), RF (random forest), AB (adaptive boosting), GB(Gradient boosting), LR (Linear regression), Voting, and ANN(artificial neural networks).

B. DCT-K

DCT-K approach begins by transforming original datasets per period into sparse representations using DCT. To illustrate our mechanism in detail, let us consider $\hat{x}$, which expresses the transformed components of the original datasets (x) after DCT [16]. To model the correlation between the transformed coefficients and energy (or information) represented among them, let us further define $\hat{x}_t$, which denotes transformed components (i.e., block) at compressed period size of N at given period t : $\hat{x}_t = \{\hat{x}_{t,1}, \hat{x}_{t,2}, \dots, \hat{x}_{t,N}\}$. Thus, each coefficient component has energy coefficient defined as: $e(\hat{x}_{t,i})$.

$EC(\hat{x}_{t,k})$ is formulated as the energy concentration (EC) contained in the number of coefficients components, denoted as K , of the entire transformed components ($\hat{x}_t$), which is calculated as:

$$EC(\hat{x}_{t,k}) = \frac{\sum_{n=1}^k e(\hat{x}_{t,n})^2}{\sum_{n=1}^N e(\hat{x}_{t,n})^2}, n = 1, 2, \dots, N, k \leq N. \quad (3)$$

Importantly, DCT-K finds k , which determines how many coefficients are required to represent δ amount of the energy [12]. k refers to the number of dominant coefficient to represent block $\hat{x}_t$ or x_t . k

IV. EVALUATIONS

A. Setup

1) *Datasets*: In our evaluation, we use real-world datasets obtained in 2020 from one of IoT weather station we have installed 11 weather stations in Korea. Datasets include 15 types of environmental factors. Among 15 environmental data, we choose five major data categories :temperature (TP), grass temperature (GTP), Relative humidity (R-HMT), soil humidity (SH), and wind speed (WS).

Table I shows the statistical properties of the original datasets in terms of Variance, skewness, and kurtosis after normalize. Skewness is a measure of data asymmetry around the mean value. Negative skewness means that more data are scattered to the left of the mean, whereas positive skewness means the opposite. The normal distribution, where data is symmetric about its mean, gives zero skewness. Measures of kurtosis in Table I indicate how outlier-prone a distribution is. As the kurtosis of any normal distribution is 3, distributions with kurtosis higher than 3 are more outlier prone. As shown in Table I, S-HMDT and WS have the high kurtosis value among all datasets, which means it is more outlier prone. In the case of skewness, R-HMDT and S-HMDT and eddy only have negative values. In the case of diff-std, R-HMDT, S-HMDT, and WS have a higher value than others. Consequently, the data characteristics of Table I gives impact on the data compression through k .

B. Performance Metrics

We want to find the data characteristics that affect lossy compression. To characterize the compression ratio of lossy compression algorithm on datasets precisely, we define the several criteria

Let $x = \{x_1, x_2, x_3, \dots, x_n\}$ be the first raw datasets, N be the number of data points in raw datasets (x), and n be the number of blocks in $\frac{\#x}{\text{blocksize}}$.

We choose the following metrics to assess the characterization of the datasets to be compressed

- Compression Ratio for DCT-K, where $|D|$ is the number of D , $|k'|$ is the number of k -dominate coefficients, is given by: $CR_K = \frac{|x| - |k'|}{|x|}$
- Let $x = x_1, x_2, x_3, \dots, x_n$ be the original data and $\hat{x} = \hat{x}_1, \hat{x}_2, \hat{x}_3, \dots, \hat{x}_n$ be the reconstructed data. Then, we measure the peak signal-to-noise ratio (PSNR), a commonly

used average error metric, especially in visualization [17], which is calculated as follows:

$$PSNR_\delta = 20 \log_{10} \left(\frac{\text{Max}(x) - \text{Min}(x)}{RMSE(x, \hat{x})} \right)$$

- Accuracy of the prediction is defined as: $\text{Accuracy} = 1 - \frac{\sum_{t=1}^n e_t^2}{\sum_{t=1}^n (y_t - y'_t)^2}$, where $e_t = (y_t - y'_t)$, y_t is real value and y'_t is prediction values. The best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse)

C. DCT-K compression

Table III shows the variation in k values using DCT-K. We use the fixed information compaction rate of 95%, 99% and 99.9%. Transforms are applied for every 60 data points, i.e, block size is 60, which are equivalent to 8,753 compression periods. As shown in Typically, periods with higher variance require more components to approximate data maintained by the same amount of energy, hence resulted in a lower approximation ratio. The WS data has relatively higher k values compared to other data sets. Therefore, the CR_K values in Table III are lower than other data sets.

D. Prediction of compression ratio

We evaluate five representative regression methods: DT, RF, AB, GB, LR, and ANN. Additionally, we add Voting algorithms against DT, AB, and RF to balance out their mechanisms.

- DT is a popular nonparametric supervised learning method to solve various real-world problems. It operates by repeatedly dividing a feature space of a given dataset using the Gini index and Entropy.
- RF is an ensemble model correlated DTs to solve the vulnerabilities mentioned above by combining multiple DTs.
- AB is also improvement of DT to fit a sequence of weak learners (i.e., models that are only slightly better than random guessing, such as small decision trees) on repeatedly modified versions of the data. The predictions from all of them are then combined through a weighted majority vote (or sum) to produce the final prediction.
- GB is a generalization of boosting to arbitrary differentiable loss functions.
- LR fits a linear model with coefficients to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation.
- Voting is to combine conceptually different machine learning regressors and return the average predicted values after DT, AB and LR.
- ANN is a mathematical model inspired by the study of the brain. It has a connected set of artificial neurons, which consists of the input layer, one or more hidden layers, and an output layer.

As previous Table III, we observed that the k increase according to δ . We predict k to represent compression ratio

TABLE I
THE CHARACTERISTICS OF EVALUATED DATASET PER COMPRESSION PERIODS (MEAN).

	k(0.95)	k(0.99)	k(0.999)	Var	Mean	Median	Skewness	Kurtosis	ADF	Kpss	dct-max	dct-std	diff-std
TP	3.21	10.39	27.47	0.08	0.52	0.51	0.03	-0.49	0.51	0.05	4.01	0.58	0.09
GTP	2.72	8.16	22.99	0.08	0.5	0.5	0.04	-0.6	0.54	0.05	3.89	0.57	0.08
R-HMDT	3.91	12.91	30.27	0.07	0.91	0.92	-0.02	-0.35	0.45	0.06	7.07	0.96	0.11
S-HMDT	9.2	18.28	28.34	0.08	9.2	9.2	-0.07	3.7	0.35	0.05	71.28	9.2	0.2
WS	12.87	26.6	39.27	0.05	0.31	0.28	0.66	1.11	0.14	0.09	2.41	0.37	0.18

TABLE II
THE EVALUATED DATASET TYPE AND COMPRESSION RATIOS IN DCT-K.

	$\delta(0.95)$	$\delta(0.99)$	$\delta(0.999)$
TP	94.6%	82.7%	54.2%
GTP	95.5%	86.4%	61.7%
R-HMDT	93.5%	78.4%	49.3%
S-HMDT	76.2%	52.7%	26.6%
WS	75%	48.3%	23.5%

TABLE III
THE EVALUATED DATASET TYPE AND ERROR RATE (PSNR) IN DCT-K.

	$\delta(0.95)$	$\delta(0.99)$	$\delta(0.999)$
TP	40.53	42.03	46
GTP	40.43	41.87	46.9
R-HMDT	31.674	33.07	37.75-
S-HMDT	49.1	51.19	54.23
WS	30.76	37.32	47.49

using the characteristics of data. We utilize machine learning libraries from scikit-learn [18] to train and evaluate five machine learning algorithms: DT, RF, AB, GB, LR, Voting and ANN. We set 70% of the data for training and the remaining 30% for the test. Figure 2 shows the performance of prediction on compression ratio for machine learning model. We evaluate performance according to fixed energy (δ : 0.95, 0.99, and 0.999). As we can see RF, GB, Voting and ANN shows better performances than others in terms of Accuracy. As we can see RF, GB, Voting, and ANN show better performances than others in terms of Accuracy. In the case of Voting, it returns the mean value of three machine model cases (LR, AB, and RF in our evaluation). The observed data characteristics in conjunction with the compression ratios show a clear correlation between those two. For example, the CR for the wind speed using 0.99 is 1.6 times higher than that using 0.999.

V. CONCLUSIONS

Our paper is to analyze recent transform-based lossy compression techniques concerning IoT datasets. This paper predicts the compression ratio of transformation-based lossy compressors, DCT-K, using characterization data to predict as statistically. Experiments show the higher the data variance, the lower the compression ratio. Especially, compression periods in DCT-K with higher data variance require more components to approximate data maintained by the same amount of energy; hence, it leads to a lower compression ratio. We evaluate machine learning to predict and observe RF, GB, Voting, and ANN

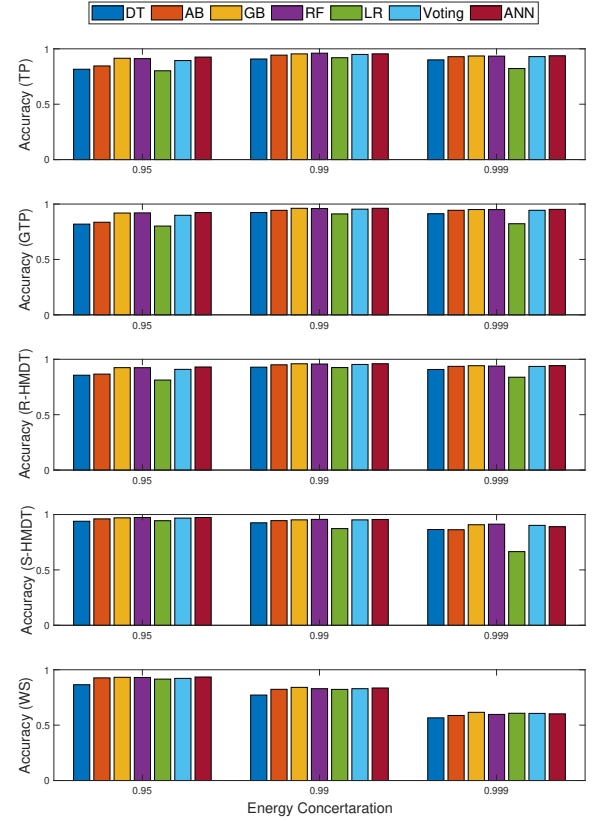


Fig. 2. Prediction Performance (Accuracy).

ACKNOWLEDGMENT

This work is supported by the Korea Innovation Foundation (INNOPOLIS) grant funded by the Korea government (MSIT) (2020-DD-UP-0278).

REFERENCES

- [1] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications," *IEEE Communications Surveys Tutorials*, vol. 17, no. 4, pp. 2347–2376, 2015.
- [2] A. Moon, K. Y. Moon, and S. W. Son, "Microclimate-Based Predictive Weather Station Platform: A Case Study for Frost Forecast," in *2017 IEEE High Performance Extreme Computing Conference*, 2017.
- [3] J. Zhang, X. Zhuo, A. Moon, H. Liu, and S. W. Son, "Efficient Encoding and Reconstruction of HPC Datasets for Checkpoint/Restart," in *Symposium on Mass Storage Systems and Technologies (MSST)*, 2019.

- [4] A. Moon, J. Kim, J. Zhang, and S. W. Son, "Lossy Compression on IoT Big Data by Exploiting Spatiotemporal Correlation," in *2017 IEEE HPEC*, 2017, pp. 1–7.
- [5] J. Zhang, A. Moon, X. Zhuo, and S. W. Son, "Towards Improving Rate-Distortion Performance of Transform-Based Lossy Compression for HPC Datasets," in *IEEE HPEC*, 2019.
- [6] T. Bose, S. Bandyopadhyay, S. Kumar, A. Bhattacharyya, and A. Pal, "Signal Characteristics on Sensor Data Compression in IoT – An Investigation," in *13th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, June 2016, pp. 1–6.
- [7] R. Chaturvedi and Y. Yadav, "A Survey on Compression Techniques for ECG Signals," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 2, no. 9, 2013.
- [8] A. Moon, S. W. Son, J. Jung, and Y. J. Song, "Understanding Bit-Error Trade-off of Transform-based Lossy Compression on Electrocardiogram Signals," in *Proceedings of IEEE Big Data*, 2020, pp. 3494–3499.
- [9] R. Sustika and B. Sugianto, "Compressive Sensing Algorithm for Data Compression on Weather Monitoring System," *TELKOMNIKA Telecommunication, Computing, Electronics and Control*, pp. 974–980, 2016.
- [10] H.-H. Chiou, A. I.-C. Lai, and C.-L. Lei, "Prediction-capable data compression algorithms for improving transmission efficiency on distributed systems," *Proceedings 20th IEEE International Conference on Distributed Computing Systems*, 2000.
- [11] A. N. Zemliachenko, R. A. Kozhemiakin, S. K. Abramov, V. V. Lukin, B. Vozel, K. Chehdi, and K. O. Egiastian, "Prediction of compression ratio for dct-based coders with application to remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2016.
- [12] A. Moon, J. Kim, J. Zhang, and S. W. Son, "Evaluating Fidelity of Lossy Compression on Spatiotemporal Data from an IoT Enabled Smart Farm," *Computers and Electronics in Agriculture*, vol. 154, pp. 304–313, Nov. 2018.
- [13] M. Rani, S. B. Dhok, and R. B. Deshmukh, "A Systematic Review of Compressive Sensing: Concepts, Implementations and Applications," in *IEEE Access*, 2018, pp. 4875–4894.
- [14] A. Moon, J. Kim, J. Zhang, H. Liu, and S. W. Son, "Understanding the Impact of Lossy Compression on IoT Smart Farm Analytics," in *2017 IEEE Big Data*, 2017, pp. 4602–4611.
- [15] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley Series in Telecommunications and Signal Processing). Wiley-Interscience, 2006.
- [16] M. A. Razzaque, C. J. Bleakley, and S. Dobson, "Compression in wireless sensor networks: A survey and comparative evaluation," *ACM Transactions on Sensor Networks*, vol. 10, no. 1, p. 5, 2013.
- [17] D. Tao, S. Di, Z. Chen, and F. Cappello, "Significantly Improving Lossy Compression for Scientific DataSets Based on Multidimensional Prediction and Error-Controlled Quantization," in *2017 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 2017, pp. 1129–1139.
- [18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.



Aekyung Moon received the Ph.D and MS degrees in computer engineering from Yeungnam University, 1997 and 2000, respectively. Since April of 2000 he has been on researcher, ETRI, Daejeon, Korea. She was a visiting scholar at University of Massachusetts University, 2017 and 2019, respectively. Her research interests are Smart City, Smart factory, AI Robot, Smart farm, etc.



Juyoung Park has joined ETRI as a member of RD group in 2001, after he got Ph.D degree from the Chungnam National University in Korea. His research areas include IoT, multicast, QoS/QoE, WPT technologies, and Digital-Agriculture. He

also has experiences in the standardization activities in ITU-T, ISO, IEC, and JTC1.



Yun-Jeong Song received the BS and MS degrees from Kyungbook National University, Daegu, Korea, and PhD degrees in electronics engineering from Chungnam National University, Daejeon, Korea in 1987, 1990 and 2004, respectively. Since July of 1990, he has been on principal researcher, Broadband Multimedia Research Team, ETRI, Daejeon, Korea. He was a visiting scholar at Center for Telecommunication Research (CTR) in Columbia University, New York from Oct. '94 to Sep. '95. His research interests are Smart City, Smart factory, AI Robot, Smart farm, etc.