Combination of Self-organizing Map and k-means Methods of Clustering for Online Games Marketing

Shaoyong Yu,[1] Mei Yang,[1] LinHai Wei,[2*]
Jian-Shiun Hu,[1] Hsien-Wei Tseng,[1] and Teen-Hang Meen[3**]

[1]School of Mathematics and Information Engineering, Longyan University, 364012 Fujian, China
[2] Fujian Xinzhi Information Technology Co., Ltd., 364012 Fujian, China
[3]Department of Electronic Engineering, National Formosa University, Huwei 632 Yunlin, Taiwan

Data mining has been applied in many fields, such as pattern evaluation, image recognition, and data analysis. Clustering is one of the most popular methods of data mining. There are many algorithms concerning clustering, such as k-means and Farthest First in data mining fields, and adaptive resonance theory (ART) and self-organizing map (SOM) in machine learning. ART and SOM are unsupervised learning algorithms, which better determine the best numbers for clustering than only the k-means algorithm. This study is devoted to applying a combination of SOM with k-means to study the marketing of online games in Taiwan. The results show that the marketing segmentation of online games can be evaluated well by clustering the users' data obtained from any online or offline survey. The method that combines SOM with k-means has been shown in this study to provide a good evaluation of the market segmentation.

1. Introduction

Owing to the popularity of mobile networks and the increasing frequency of broadband, the Internet has come to play an important role in people's lives. In 2017, according to the "Taiwan Broadband Network Usage Survey" published by the Taiwan Network Information Center (TWNIC), the total number of Internet users in Taiwan reached 17.6 million. Because of the popularity of mobile Internet, Taiwan's Internet population is approaching the total population of the country. According to the Market Intelligence & Consulting Institute (MIC), the game players in Taiwan were surveyed in 2017.[1] Among the online game types of the digital game category, up to 75.5% of the players are most interested in mobile/tablet app games, overwhelmingly surpassing the second-place computer online games (27.5%), while other game types such as computer web games (25.5%), computer stand-alone games (21.5%), and video console games (16.5%) are the last to be played. Although virtual reality (VR) games have enjoyed high exposure in recent years and are beginning to gain a following, at 2.6%, there is still a considerable distance to go before they can join the mainstream games. The global digital

---

game market is booming and rife with huge business opportunities, providing a receptacle for new technology applications such as VR, augmented reality (AR), and 5G. With the popularity of platforms such as mainframes, computers, and mobile devices, games themselves infiltrate players' lives and drive live broadcast, advertising placement, and other industries. Taiwan has a thriving game industry with active users and live program trackers, and its development potential is endless.[2] In this study, the term online game generally refers to digital game activities through mobile, wireless, and wireline media. According to the results of the above research, the online game population accounts for a considerable proportion of the general population.

Therefore, in the highly competitive online game market, it is important for game manufacturers to understand the diverse attributes of their video game customers and the preferences of related products, and to expand the market of video games. In the marketing strategy, because of limited products and resources, it is impossible to fully expand all the test sites, effectively separate the market, and find the best market position (niche). Using an online questionnaire, we obtain the opinions and preferences of the online game customer group, and then conduct clustered data exploration (clustering). The data are divided into different attribute groups, allowing us to effectively separate customers and judge the contribution of different attribute groups. Then, by selecting the most suitable potential customers from the indicators or constraints, the manufacturer can implement appropriate marketing strategies for these potential customers. For the clustering data method, at present, machine learning is the mainstream. Among the machine learning methods are artificial neural networks (ANNs), such as the adaptive resonance theory network (ART) and self-organizing map (SOM). Also, cluster analysis in data mining, such as the k-means algorithm, has the characteristics that "the highest similarity within the group and the lowest similarity between groups" can effectively segment customers and uncover the best market position. Thus, for commercial application, cluster analysis provides a better decision making method for improving the effectiveness of customer relationship management. This study can be applied to IOT development in the market, since every customer is like a target or an item for marketing segmentation. We can develop an interface module with the ability to aggregate data collected from every customer resembling sensors that can detect the intention of a customer browsing on the Internet. In the future, by applying SOM and k-means as the core of systems, we can develop a complete IOT system for marketing segmentation.

The paper is organized as follows: we mainly introduce the SOM algorithm in machine learning in Sect. 2 and the k-means algorithm in data grouping exploration in Sect. 3. In Sect. 4, we explain the importance of market segmentation in the online game industry operation and the latest related research. The integration of SOM and k-means for online game customer segmentation is discussed in Sect. 5. Research results are discussed and research conclusions are presented in Sects. 6 and 7, respectively.

2. SOM

SOM, in the field of machine learning, is an unsupervised learning network model, which was proposed by Kohonen in 1990.[3] The basic principles of the SOM network can be traced

back to the characteristics of the brain structure; brain cells with similar functions in the brain gather together. The brain nerve cells exhibit this characteristic of "similar things come together". SOM copies this characteristic, and its output processing units affect each other. When the network is learning, its output processing units will have similar functions, that is, have similar link weights.[4]

SOM is a network architecture consisting of an input layer and an output layer. In the input layer, neurons represent the attributes of the input data, are independent of each other, and have independent weights. For the output layer, a two-dimensional grid network structure is commonly used, and the output and input-layer neurons have interconnected network relationships in a manner represented by a weight vector.

In 2000, Vesanto and Alhoniemi[5] proposed using SOM for clustering analysis to obtain better data clustering and reduce computation time. In 2005, Bacao *et al.*[6] proposed that SOM can be used instead of k-means for data clustering. In recent years, research and application in related fields have gradually confirmed that SOM and k-means can be combined to create a better method for data clustering.[7–9]

3. Clustering Analysis by k-means Method

k-means is the simplest and most commonly used algorithm employing a squared error criterion.[10] It starts with a random initial partition and keeps reassigning the patterns to clusters on the basis of the similarity between the pattern and the cluster centers until a convergence criterion is met (e.g., there is no reassignment of any pattern from one cluster to another; otherwise, the squared error ceases to decrease significantly after a number of iterations). Grouping can be divided into two major architectures: hierarchical clustering and partitional clustering, as shown in Fig. 1.

4. Marketing Segmentation in Online Games Industry

Market segmentation is the process of dividing the mass market into groups with similar needs and wants.[11] The rationale for market segmentation is that in order to gain a competitive advantage and superior performance, firms should "(1) identify segments of industry demand, (2) target specific segments of demand, and (3) develop specific 'marketing mixes' for each
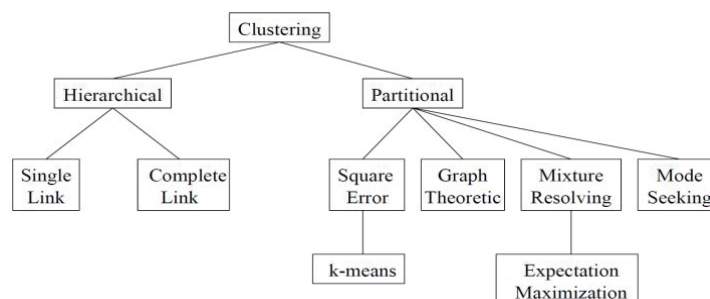

Fig. 1. Taxonomy of clustering approaches.

targeted market segment".[12]   From an economic perspective, segmentation is based on the assumption that heterogeneity in demand allows for the demand to be disaggregated into segments with distinct demand functions.[13]

Because of the boom in online games, the market is fiercely competitive, and much research on the online game industry is being carried out. The Taiwanese game industry mostly fell into hard times, such that it is still important to understand the characteristics of players and the utmost importance of marketing. From the responses to online questionnaires regarding the preferences of online game players, Lee elucidated the decision making process of choosing games.[14] The results of two-stage grouping show that there is a large difference between the ideas of the players and online game companies.[14]   The online game market characteristics found through research can be used as a reference for marketing. Through an online game industry survey, Ren and Hardwick[14] found that the 11 online game makers in Taiwan, for example, used the number of online game strategies they evaluated to divide the industry into (1) leaders, (2) challengers, (3) followers, and three other types, as a business strategy reference for similar operators.

5.   Combining SOM with k-means for Marketing Segmentation

5.1   Proposed model for marketing segmentation

Effective marketing segmentation is an important marketing strategy for online game business operators. In market segmentation, questionnaire surveys were used in the past, and the responses were analyzed by factor analysis and related multivariate analysis in statistics. In the era of big data, the amount of data collected through the Internet is large, and machine learning and data exploration have begun to be used for analysis and application. Kuo and Liao[15] proposed a clustering model based on ART combined with k-means to determine online game players' attributes; the cluster number A obtained by ART and the group number B obtained by k-means were found to coincide. Common attribute groups could be identified, providing vendors with a means of better customer relationship management.

On the basis of the above literature, although the k-means algorithm is the simplest and most commonly used clustering method, if the number of clusters is not specified in advance or the software is randomly selected, clustering may not be possible. Furthermore, the accuracy of clustering depends mainly on the researcher's experience in appropriately applying the algorithms. The SOM-like neural model is more accurate than the traditional multivariate method.[16] Zhang and Tsay[17] proposed a two-stage strategy clustering method to improve the accuracy of grouping. In the first stage, SOM is used to map high-dimensional data to a low-dimensional space to easily understand the data structure. With a parameter S of rough groups, convergence can achieve global optimization by using the software "Viscovery SOMine". In the second stage, the k-means algorithm of Weka software is used for cluster analysis by setting the number of K clusterings to be equal to S. The purpose is to find the characteristic pattern of the data so that the difference between the attributes in the same groups is as small as possible. We propose to integrate the SOM and k-means methods into the SOM-k architecture, as shown in Fig. 2.
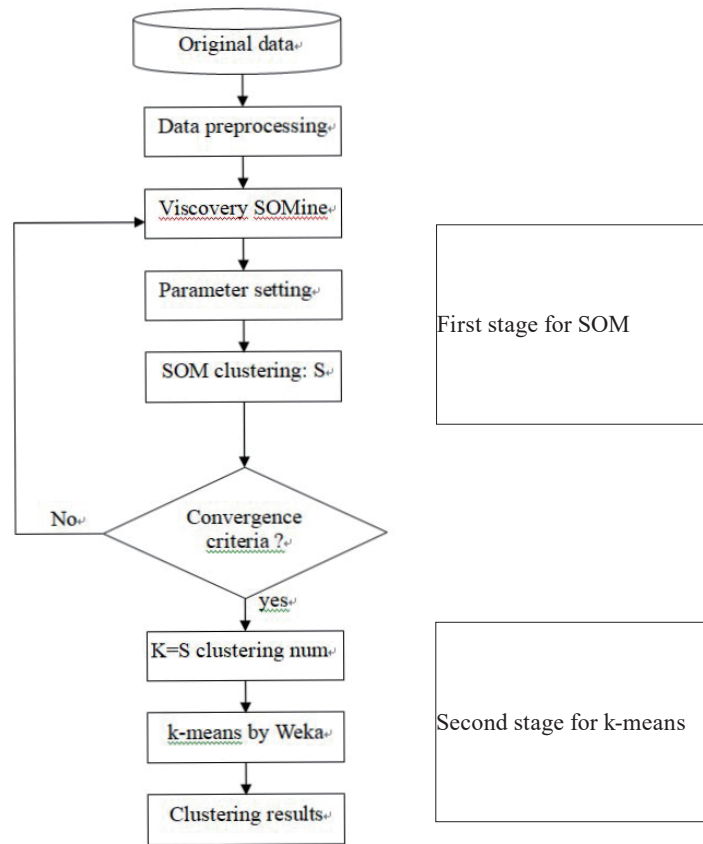
Fig. 2.   SOM-k clustering process.

5.2   Data description for SOM-k

The data for online games marketing was collected by both online and offline surveys in 2011. There were 473 responses received, 438 of which were effective data. The attributes of the data are listed in Table 1.

Table 1
Description of data attributes.

Attribute	Description
Age	Below 12, 13–18, 19–24, 25–29, Above 30
Occupation	Student, Services, IT, Manufacture, Unemployed, Others
Monthly expense (NTD)	Below 100, 101–300, 301–500, 500–700, Above 700 NTD
Education	Junior high school, Senior high school, College, Bachelor, Master
Online game (played for at least 6 months)	Lineage, Ragnarok Online 2 (RO2), World of Warcraft, Maple Story, Mabinogi, Lineage 2, New Hope, Cross Gate

5.3   Data preprocessing

The attribute data in this study are the numerical and categorical data types, but the SOM clustering algorithm can only process numerical data. However, using the co-occurrence

clustering method, Zhang and Tsay[17] considered the accompanying relationship between the categories of attributes in the data set and numerically classified all the category attributes in accordance with this relationship. Then, mixed data were converted to full-value data. In this manner, grouping can be accomplished by various clustering methods, and data of this type are used for numerical conversion. As for the preprocessing of the numerical data, normalization should be performed before running the SOM algorithm, in order to obtain a better and more suitable equation for reliable clustering outcomes. The formula of normalization is

$$\overline{x}_i = \frac{x_i - x_i^{min}}{x_i^{max} - x_i^{min}}, \tag{1}$$

where x_i^{max} represents the maximum value of the original data in the attribute, x_i^{min} represents the minimum value of the original data in each facet, and i is the value of each attribute category, for example, for the original data attributes {A, B, C}, i = A, B, C. The normalized data of the attributes "Age" and "Monthly expense" are shown in Tables 2 and 3, respectively.

Second, the categorical data type can be transformed into the numerical data type by the co-occurrence clustering method. This is a complex and tedious procedure.[18] However, the group of mixed data to determine relationships between attributes is similar to the distance between the data objects. After the above procedure is completed, the category attribute is digitized as shown in Table 4–6.

Table 2
Normalized numerical attribute "Age".

Age	Mean	Normalized
[0,12]	6	0
[13,18]	16	0.38
[19,24]	22	0.62
[25,29]	27	0.81
[30,∞]	32	1

Table 3
Normalized numerical attribute "Monthly expense".

Monthly expense	Mean	Normalized
[0,100]	50	0
[101,300]	200	0.2
[301,500]	400	0.47
[501,700]	600	0.73
[700,∞]	800	1

Table 4
Numerical values for online games.

Online game	Numerical value
World of Warcraft	0.6618
Lineage	0.6455
Ragnarok Online 2 (RO2)	0.5593
Maple Story	0.5156
Mabinogi	0.4614
Lineage 2	0.7646
New Hope	0.5839
Cross Gate	0.5589

Table 5
Numerical values for occupation.

Occupation	Numerical value
Student	0.5562
Services	0.6094
IT	0.6466
Manufacturing	0.6321
Free	0.6122
Others	0.6067

Table 6
Numerical values for education.

Education	Numerical value
Junior high school	0.5769
Senior high school	0.5881
Bachelor	0.5814
Master	0.5919
College	0.616

5.4 SOM-k exploration results

After the above data preprocessing steps, the 438 samples are converted into SOM-k for data cluster exploration. Of the parameters calculated by SOM, map size is the number of generated nodes. The more nodes, the more detailed the map, and the topology nodes are preset in hexagonal order. The smaller the average of the attributes, the more data space is required, typically between 0.5 and 2 normalized unit. The training procedure is under a good convergent rate. The calculation results show that the best cluster number is 3. The network topology space is shown in Fig. 3.

After SOM determines the optimal cluster number, the cluster number is set to the cluster number of k-means, and clustering is performed to obtain the final clustering result. Table 7 shows the centers of the average attributed parameters of each of the three clusters.

6. Results and Discussion

The results of the SOM-k cluster exploration are summarized in Table 8.

The results of the analysis of each of the three groups are most significant in terms of age, and the following three groups can be obtained.
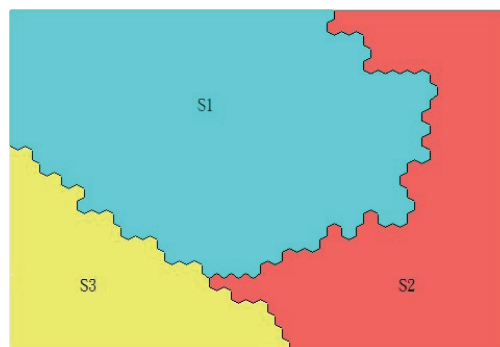

Fig. 3.   (Color online) SOM network topology space grouping diagram.

Table 7
Geometric centers of each cluster's attributes.

	Age	Education	Occupation	Monthly expense	Online game	Size (%)
S1	0.807	0.3305	0.3523	0.490	0.3492	44.10
S2	0.553	0.3300	0.3121	0.167	0.3139	38.92
S3	0.778	0.3473	0.3468	0.409	0.3480	16.98

Table 8
Results of SOM-k clustering.

Clustering size (%)	Age	Education	Occupation	Monthly expense (NTD)	Online game (long term)
54	19–24	College	Student	Below 100	Ragnarok Online 2 (RO2)
31	25–29	College	Services	300–500	World of Warcraft
16	13–18	Senior high school	Student	Below 100	Maple Story

Table 9
Players' ratings for each game.

Ranking	Game name	Score
1	World of Warcraft	85.2767
2	RO	79.2436
3	Lineage 2	78.1650
4	Lineage	76.4706
5	Mabinogi	75.7531
6	Cross Gate	75.6939
7	Maple Story	74.6815
8	New Hope	74.4510

Group 1: University students: age 19 to 24 years, students, prefer "RO2".

Group 2: Social youth group: new members of workforce, college level, age 19 to 24 years, prefer "World of Warcraft".

Group 3: Junior high/high school students: age 13 to 18 years, students, prefer "Maple Story".

The results of this study were compared with those for experienced players in 2011[18] shown in Table 9.

The social youth group prefers World of Warcraft, which experienced players ranked first. The new Ragnarok online legend, which was preferred by the students group, is second in the ranking of experienced players. The junior high/high school students group prefers Maple Story, ranked 7th by experienced players. Since games in this ranking were evaluated by experienced players, Maple Story is judged to be simple and easy to operate, and its age groups are biased towards younger players. In this study, the majority of players were 19–29 years old; thus, it is foreseeable that the ranking results of games played by this group will not be high. According to the above description, the cluster number of this study is 3. Two clusters are associated with games ranking first and second in the experienced players' evaluation of games, which verifies the accuracy of the grouping in this study relative to the popularity of actual online game products. The current market situation is consistent with our results.

7. Conclusions

In this study, we investigated the SOM network and the two-stage grouping method supplemented by the k-means algorithm. The number of clusters of online game players was determined and named as junior high/high school students, university specialists, and social youth groups. The ranking of games as evaluated by experienced players indicated that the clustering results are reliable. After we group the data, we can adjust or refer to the decision based on the clustering result. Suppose there is a manufacturer developing a new online game. The game type is similar to Maple Story. To attract more potential customers, the manufacturer can promote the game mainly through TV shows or popular websites that are often watched by high school students, who are the main targets. In contrast, if game makers choose professional forums or advertise on the financial channel, the relative publicity benefits will be greatly reduced.

According to the research results, manufacturers can realize better customer management and effective promotion in appropriate places, while reducing the time and resources spent in inappropriate places, thus avoiding the waste of resources. This is part of every company's publicity plans. From the groupings of clusters, we also learned that the current World of Warcraft and new Fairyland Legends games can be found among the current diversified online games, and their popularity has always been among the highest on game discussion boards. These two types of online games are still the ones most players like. Game makers can design new games on the basis of a complete analysis of their architecture or popular attributes.

In this study, the characteristics of players in large-scale online games were examined. Time and financial constraints prevent us from exploring media such as Facebook in future research and comparison of online games. Since online game players are mostly students, if the number of samples can be increased in the future, the player characteristics in the game market will be better represented.

In the future, by using the following concept, we can develop a complete IOT system for marketing segmentation, which can be constructed for future development in the marketing field. Since every customer is a target or an item for marketing segmentation, those modules, which have the ability to aggregate data collected from every customer, resemble sensors for detecting the intentions of customers browsing on the Internet. All of the decision making in marketing segmentation can be automated in the web systems by using the virtual sensors throughout all the browsing on the web.

Acknowledgments

References

1 A Survey on Broadband Internet Usage in Taiwan: https://www.twnic.net.tw/doc/twrp/20170721d.pdf (accessed January 2020), Taiwan Network Information Center.
2 The Taiwanese Digital Gaming Market Overview: https://www.giiresearch.com/report/iii668781-taiwanese-digital-gaming-market-overview.html (accessed January 2020).
3 T. Kohonen: Proc. IEEE **78** (1990) 1464. https://doi.org/10.1109/5.58325
4 S. Benzer and R. Benzer: Enterprise & Business Management (Tectum Wissenschaftsverlag, Marburg, 2020) pp. 287–310.
5 J. Vesanto and E. Alhoniemi: IEEE Trans. Neural Networks **11** (2000) 586.
6 F. Bacao, V. Lobo, and M. Painho: Proc. Computational Science – ICCS 2005 (Springer, 2005) 476–483.

7   S. S. Abidi and J. Ong: 2000 TENCON Proc. Intelligent Systems and Technologies for the New Millennium (IEEE, 2000) 568–573.

8   C. H. Hung and C. F. Tsai: Expert Syst. Appl. **34** (2008) 780.

9   G. Zhu and X. Zhu: 2010 Int. Conf. Artificial Intelligence and Computational Intelligence (IEEE, 2010) 178–182.

10   J. Mcqueen: Proc. 5th Berkeley Symp. Mathematical Statistics and Probability (1967) 281.

11   W. M. Pride, O. C. Ferrell, B. A. Lukas, S. Schembri, O. Niininen, and R. Cassidy: Marketing Principles (Cengage Learning, Boston, 2018) 3rd Asia-Pacific ed., p. 638.

12   S. Madhavaram and S. D. Hunt: J. Academy Marketing Sci. **36** (2008) 67.

13   P. R. Dickson and J. L. Ginter: J. Marketing **51** (1987) 1. https://doi.org/10.2307/1251125

14   Q. Ren and P. Hardwick: Electronic Globalized Business and Sustainable Development Through IT Management: Strategies and Perspectives (Business Science Reference, Hershey, 2010) p. 135. https://doi.org/10.4018/978-1-61520-623-0.ch010

15   R. J. Kuo and C. L. Liao: J. Chinese Inst. Industrial Eng. **19** (2002) 64. https://doi.org/10.1080/10170660209509214

16   H. Y. Zhao and S. X. Zhao: Proc. 2012 Int. Conf. Information Technology and Management Science (ICITMS 2012) (2013) 55–60.

17   H. H. Zhang and S. F. Tsay: J. Inf. Manage. **11** (2004) 161.

18   Y. Zhang and W. Huang: MATEC Web of Conf. **267** (2019) 1. https://doi.org/10.1051/matecconf/201926704010

About the Authors

Shaoyong Yu received his B.S., M.S., and Ph.D. degrees from Xiamen University, China, in 2004, 2007, and 2017, respectively. From 2007 to 2017, he was an assistant professor at Xiamen University of Technology, China. Since 2018, he has been a professor at Longyan University, China. His research interests are in machine learning, computer vision, and pattern recognition. (82018007@lyun.edu.cn)

Mei Yang received her B.S. degree from Fujian Agriculture and Forestry University, China, in 2009. Since 2019, she has been an associate professor at Longyan University, China. Her research interests are in machine learning and computer vision. (92019002@lyun.edu.cn)

LinHai Wei received his B.S. degree from Renmin University of China, in 2014. He founded a company in 2011 and served as a chairman and a manager. He obtained multiple qualification certifications and declared three invention patents. (85567415@qq.con)

Jian-Shiun Hu received his B.Eng. degree from National Taiwan University, in 1983 and his M.Eng. and Ph.D. degrees from National Cheng Kung University, Taiwan, P.R. China in 1989 and 1994, respectively. From 1994 to 1996, he was an associate researcher in the High Performance Computing Center, National Science Council of Taiwan. From 1996 to 2000 and 2000 to 2018, he was an associate professor in MingShin University of Science and Technology and Hsuan Chung University, Taiwan, P.R. China, respectively. He is currently a professor in the School of Mathematics and Information Engineering, Longyan University. His current research interests include machine learning, big data, neural networks, and data mining.

Hsien-Wei Tseng received his B.S. degree in electrical engineering from National Kaohsiung University of Applied Sciences in 2002 and his Ph.D. degree in electrical engineering from Tamkang University in 2010. He joined the School of Mathematics and Information Engineering, Longyan University, Longyan, Fujian, China, in Aug. 2017 and now serves as a professor. His main research interests include system design and performance evaluation in IoT and wireless communication systems.

Teen-Hang Meen received his B.S. degree from the Department of Electrical Engineering, National Cheng Kung University (NCKU), Tainan, Taiwan in 1989, and his M.S. and Ph.D. degrees from the Institute of Electrical Engineering, National Sun Yat-Sen University (NSYSU), Kaohsiung, Taiwan in 1991 and 1994, respectively. He was the chairman of the Department of Electronic Engineering of National Formosa University, Yunlin, Taiwan from 2005 to 2011. He received the excellent research award given by National Formosa University in 2008 and 2014. Currently, he is a distinguished professor with the Department of Electronic Engineering, National Formosa University, Yunlin, Taiwan. He is also the president of the International Institute of Knowledge Innovation and Invention (IIKII) and the chair of the IEEE Tainan Section Sensors Council. In recent years, he has published more than 100 SCI and SSCI papers.