

Deep Learning and Graph-Based Indexing for Context-Aware Video Retrieval Systems

Dinh-Lam Pham¹, Lanh Thanh Le², Jeong-Hyun Chang¹, Sang-eun Ahn¹, Kyonghee Sun¹,
Kyong-Sook Kim¹, and Kwanghoon Pio Kim^{1*}

¹Contents Convergence Software Research Institute, Kyonggi University, South Korea

{phamdinhlam, crime_tiger564, sangeun1804,
sunkh0507, khmjmc, kwang}@kgu.ac.kr

²Faculty of Information Technology, Dong Nai Technology University, Bien Hoa City, Vietnam

lethanhlanh@dnctu.edu.vn

Abstract— Managing the CCTV recorded footage takes advantage of not only hardware resources but also labor costs in monitoring and searching content through vast amounts of video clips generated every minute from CCTV cameras. Many studies and practices were introduced to deal with this matter. However, several challenges still need to be solved when implementing a video retrieval system. This study explores the applications of deep learning and graphs to create indexed data in order to implement functions in a video retrieval system. In particular, in this paper, we proposed an approach using graph structure to index the content in CCTV videos after using deep learning algorithms to contextualize detected objects in frames from these videos. Furthermore, we extend the approach to manage all the CCTV videos using graph management approach. This strategy, on the one hand, takes advantage of graph structure, providing an intuitive view for the CCTV management centers in managing and searching content from the managed cameras. On the other hand, it offers easy and efficient retrieval of video content.

Keywords— Graph-based indexing, Video contextualization, Deep learning-based system, Contextualization dataset, Video retrieval system, CCTV management system.

I. INTRODUCTION

Nowadays, CCTV cameras have become popular and appear everywhere to support managing security in public and private areas. CCTV is used to monitor the activities of citizens in public places such as shopping malls, streets, and stations or to patrol the security of sensitive and high-risk locations like government buildings, banks, etc. Recently, more and more organizations have used CCTV systems to enhance security and ensure safety for their entities. Furthermore, with the growing worry of crime, CCTV systems are essential tools supporting police and government departments in managing social stability. Subsequently, CCTV cameras play a significant role in supporting and controlling order in public areas [1][2][3].

Using the recorded videos from CCTV cameras, police officers can find critical information from footage in frames and then support investigations, identify and follow the suspects, or track abnormal activities. However, searching and finding the content in recorded video files from CCTV cameras is not an

easy task. Since CCTV cameras record scenes every second and there are hundreds of cameras in each area, a huge number of videos are recorded that need to be watched by the naked eye to search and find the content in each camera.

Beside the advantage, there are several challenges in searching and retrieving content from footage generated from CCTV cameras. The first and primary major issue is the large size of video files, which consumes time and makes it difficult to manually search over hours of recorded content. It is hard work to manually seek through hours and days of videos and not miss any important frames. These tasks become even more difficult with the systems that lack effective indexing or tagging functions in the recorded content [4][5]. In these cases, to search content in video files, police officers have to review the video frame by frame, and it is inevitable that this process will become increasingly inefficient and time-consuming.

For helping video retrieval systems, many approaches were introduced in order to create efficient and fast functions in searching content from many video files. For example, using automatically segmenting movies into logical story units supporting advanced video retrieval systems [6], mining temporal patterns from video contents with efficient indexing method and matching techniques [7], using Apache Hadoop and Storm supporting retrieval system [8], using Fuzzy-based SVM classification method [9], using Segmentation mechanisms [10], using Multi/cross-model strategies [11] [12] [13][14], using the Vision transformer and CLIP [15] [16], and some other methods.

Even though several methods were proposed and introduced in supporting video retrieval systems, there is still more work to be done to create more and more efficient video retrieval systems. In this paper, we propose an approach using deep learning and graph-based indexing mechanisms to assist video retrieval system by visualizing the overall picture of the content recorded in the video and hence supporting retrieval content inside it. More precisely, in the strategy, we:

- Contextualize video content using the contextualization mechanism introduced in [17].
- Index the detected content in a video and then create a map indexing all videos using an undirected graph structure.
- Represent the indexed data with graph visualization and hence search video content through the graph and indexed data.

The proposed method provides a comprehensive view of video content recorded from CCTV cameras as well as provides easier video retrieval function.

The remaining content of this paper is categorized as follows: The next section discusses some recent works related to video retrieval and image content searching methodologies. In section III, we describe our proposal in detail, about the concept and procedure to generate the graph-based indexing content supporting retrieval function. Finally, we present some experimental results and conclusions of our proposal.

II. RELATED WORKS

This section discusses the related research about the topic of video retrieval and image content searching for giving us the overall picture of methodologies in this topic.

In [18] Su et al. introduced an approach using pattern-indexing and matching techniques to estimate the optimal outcome in retrieving content from video. In particular, the method contained three main stages, i.e., Preprocessing, Indexing, and Searching. In the Preprocessing stage, they processed the query clip and target video by dividing a video into a set of sequential shots (shot detection) and then extracting the features using color layout and edge histogram from the shot. Finally, the clustering (k-means) and encoding were used to create different encoded-clusters. Later, the Indexing stage generated temporal patterns and constructed the index-Tree of the input query clip and videos to generate FPI-tree (fast-pattern-index tree) and AFPI-tree (Advanced Fast-Pattern-index tree). In the final Search stage, they compared the matching patterns and re-ranking results to get the Top-k results of the input clip. By focusing on the temporal patterns, the proposed method improves retrieval speed and accuracy as well as reduces the complexity of visual feature matching. Even so, because of relying on visual features (color and edge histograms), it makes challenging to distinguish between similar categories. Furthermore, the construction of trees (FPI and AFPI) can require many computer resources.

In [19] Dong et al. proposed a method using a dual encoding neural network for retrieving video content by text. The concept of the proposed network is that it performed in parallel multi-level encoding (i.e., mean pooling, biGRU, biGRU-CNN) for both the input video and text. After that, hybrid space learning was used in combined vectors for each input. These vectors are projected into the space with latent and concept. After the

training face, the encodings for video and text can be processed independently to allow the large-scale videos to be processed offline; meanwhile, ad-hoc queries will be answered in real-time. The proposed method improved the quality of the feature by capturing not only global but local and temporal information as well as including both video and text. The approach is scalable for large-scale retrieval tasks, and input videos can be processed offline. However, the dual encoding network's complexity increases as the method uses multi-level encoding and the hybrid space for learning both the input data from video and text. It is inevitable that more computational resources and time are needed for training the network.

In [20] Zhao et al. proposed CenterCLIP, a token clustering method for efficient text-video retrieval system. In the method, they introduced a multi-segment token clustering approach to improve the efficiency of retrieval with text-video. By reducing the number of redundant tokens of continuous frames and achieving segment-level alignment, the efficiency of retrieving function was significantly improved. In particular, the procedure of the method is as follows. The input video is divided into multiple temporal segments, each frame within a segment is tokenized and hence producing a sequence of visual tokens. K-medoids++ and Spectral clustering were used to select the representative tokens from each video segment. After that, the selected tokens from all segments are concatenated into a new token sequence and processed through the transformer-based attention algorithm, giving the model ability to learn features from video. Finally, the CLIP model is used to encode both the video and text data into a shared embedding space to learn high-quality representation of both frames and input text. The CenterCLIP achieved state-of-the-art performance on the multiple video retrieval benchmarks.

In our study, we take advantage of graph structure to organize and index the detected objects in video frames using deep learning methods, hence supporting the retrieval of video content.

III. PROPOSED METHOD

A. Context-aware Video Retrieval Content

In this section, we explain the concept of context aware video retrieval and our proposed method for retrieving video content using graph-based indexing mechanism.

- Let V be a sequence of frames in the video:
$$V = \{f_1, f_2, f_3, \dots, f_N\}$$
- Let O be the set of all possible object classes detected in the video frames
$$O = \{o_1, o_2, o_3, \dots, o_M\}$$
- For each frame f_i , Let C_i be a vector representing the counts of each object class in that frame:
$$C_i = [c_i^{o_1}, c_i^{o_2}, c_i^{o_3}, \dots, c_i^{o_M}]$$
- Let Q be a set of searching input where the input is on the count of a particular object class

$$Q = \{(o, R_o, k) \mid o \in O, R_o \in \{=, <, >, \leq, \geq, \neq\}, k \in N_0\}$$

The context-aware video retrieval content is the set $\mathcal{R}$ of frames that satisfy the input query Q :

$$\mathcal{R} = \{f_i \in V \mid \forall (o, R_o, k) \in Q, c_i^o R_o k\}$$

For example: Search the frames in the video containing more than k_1 cars where the number of persons is larger than k_2

$$Q = \{(car, >, k_1), (person, >, k_2)\} \\ \Rightarrow \mathcal{R} = \{f_i \in V \mid (c_i^{car} > k_1) \text{ and } (c_i^{person} > k_2)\}$$

The context-aware content $\mathcal{R}$ returns a list of frames containing the number of detected cars and persons larger than k_1 and k_2 .

B. Video Objects Contextualization Machine (VOCM)

For preparing video contextualization data to be indexed and used in the retrieval phase, we create a video contextualization machine to automatically generate the JSV (JSON video content) files containing the detected objects in each frame from the input videos. Figure 1 illustrates the operation of the VOCM.

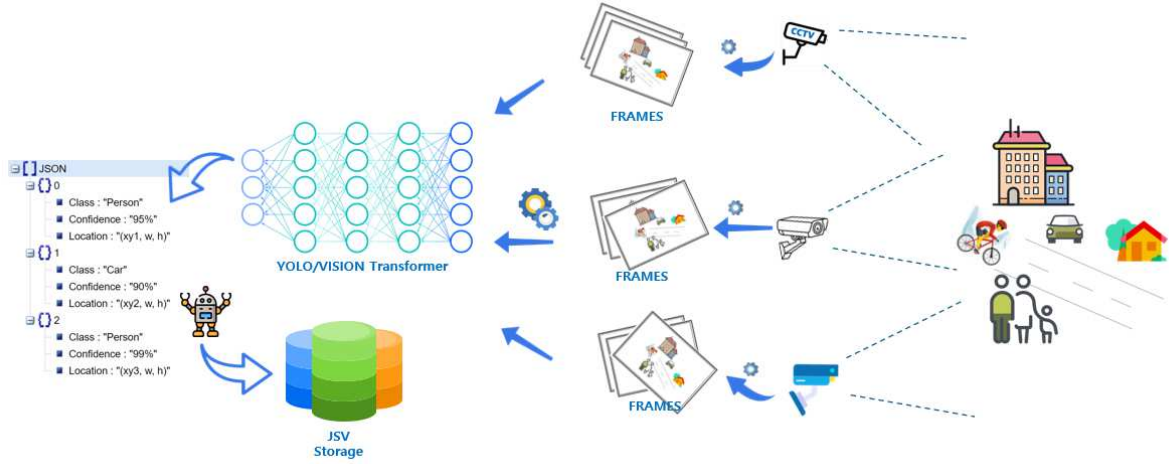


Figure 1. The concept of Video Objects Contextualization Machine

C. Graph-based Indexing Mechanism

Every minute, CCTV cameras record and store vast amounts of video content. Efficiently searching and figuring out a specific scenario from the storage videos is a big challenge. To address this concern, we propose the Graph-based indexing mechanism for creating a map and overall picture of the content of a group of recorded videos. To be more precise, we introduce a graph-based method to manage the videos through detected content in JSV files.

The idea of the graph-based indexing method is that. For managing a group of videos, let's suppose each video is a node in a graph; the detected objects in each video are the other individual nodes. An edge connecting the video and the object represents the relationship between the video and detected objects. Similarly, suppose each frame is a node for managing and retrieving content in a video; the detected objects in each frame are the other individual nodes in the graph. The relationship between the frame and detected objects is represented by an edge. Figure 2 represents the algorithm of the proposed method.

Algorithm 1 Graph-Based Indexing Algorithm

```

1: Load the JSV file
2: Initialize an empty graph  $G$ 
3: for each video  $V$  in JSV["videos"] do
4:   Add a node to  $G$  representing the video
5:   for each frame  $F$  in  $V$ ["frames"] do
6:     Add a node to  $G$  representing the frame
7:     Create an edge between the current frame and the video
8:     for each detected object  $O$  in  $F$ ["objects"] do
9:       Extract object class, location and other properties from  $O$ 
10:      Check if a node for object class  $O$ ["class"] already exists in  $G$ 
11:      if not then
12:        Add a node to  $G$  representing the object class
13:      end if
14:      Create an edge between this node and its belong frame
15:      Store the information of node as an attribute of the edge and node
16:    end for
17:  end for
18: end for
19: return  $G$ 

```

Figure 2. Graph-based Indexing Algorithm

IV. EXPERIMENTAL RESULTS

To evaluate the proposed strategy, we conduct installing the system in a CCTV management center and record video in one week. Figure 3 demonstrates the results of the implemented system for managing and retrieving video content from CCTV videos. As we can see on the figure, using graph-based indexing

strategy provided a comprehensive view of the content of the video, supported managing and retrieving the video content intuitively.

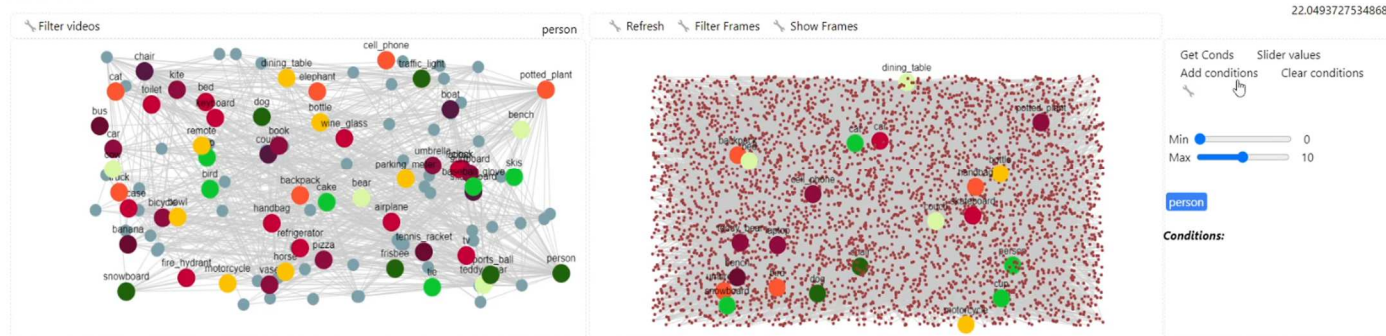


Figure 3. Managing and Retrieving Video Content

V. CONCLUSIONS

In this study, we proposed and introduced a strategy to manage and retrieve video content using contextualization data with graph-based mechanism. We:

- Proposed the definition and concept of context-aware video retrieval content, video object contextualization machine, and graph-based indexing mechanism.
- Presented the method and algorithm to automatically contextualize video data to JSV and index JSV files.
- Evaluated the proposed method by implementing a system with CCTV cameras.

Beside the achievement from this proposal, the proposed method still has some limitations and challenges to be addressed and solved in the future research for fully supporting video retrieval system, such as adding more condition into input searching queries, deal with graph analysis, etc. In future work, we plan to improve the system by making it faster and more accurate, ensure it can handle different types of inputting data and keep privacy of video content.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (grant number 2020R1A6A1A03040583). This work was also supported by National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT, Ministry of Science and ICT), Republic of Korea (Grant No. NRF-2022R1A2C2093002).

REFERENCES

- [1] M. P. J. Ashby, "The Value of CCTV Surveillance Cameras as an Investigative Tool: An Empirical Analysis," *Eur. J. Crim. Policy Res.*, vol. 23, no. 3, pp. 441–459, Sep. 2017, doi: 10.1007/S10610-017-9341-6.
- [2] E. L. Piza, J. M. Caplan, and L. W. Kennedy, "Analyzing the Influence of Micro-Level Factors on CCTV Camera Effect," *J. Quant. Criminol.*, vol. 30, no. 2, pp. 237–264, 2014, doi: 10.1007/S10940-013-9202-5.
- [3] M. H.-C. & P. Pol'y and undefined 2019, "Security or surveillance: Examination of CCTV camera usage in the 21st century," *HeinOnline*, Accessed: Sep. 12, 2024. [Online]. Available: https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/crpp18§ion=15
- [4] N. Spolaôr, H. D. Lee, W. S. R. Takaki, L. A. Ensina, C. S. R. Coy, and F. C. Wu, "A systematic review on content-based video retrieval," *Eng. Appl. Artif. Intell.*, vol. 90, p. 103557, Apr. 2020, doi: 10.1016/J.ENGAPPAL.2020.103557.
- [5] N. Spolaôr, H. D. Lee, W. S. R. Takaki, L. A. Ensina, C. S. R. Coy, and F. C. Wu, "A systematic review on content-based video retrieval," *Eng. Appl. Artif. Intell.*, vol. 90, p. 103557, Apr. 2020, doi: 10.1016/j.engappai.2020.103557.
- [6] A. Hanjalic, R. L. Lagendijk, and J. Biemond, "Automated high-level movie segmentation for advanced video-retrieval systems," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 4, pp. 580–588, 1999, doi: 10.1109/76.767124.
- [7] J. H. Su, Y. T. Huang, and V. S. Tseng, "Efficient content-based video retrieval by mining temporal patterns," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 36–42, 2008, doi: 10.1145/1509212.1509217.
- [8] E. M. Saoudi, A. A. El Ouadrhiri, S. J. Andaloussi, O. El Warrak, and A. Sekkaki, "Content based video retrieval by using distributed real-time system based on storm," *Int. J. Embed. Real-Time Commun. Syst. (IJERTCS)*, vol. 10, no. 4, pp. 60–80, Oct. 2019, doi: 10.4018/ijertcs.2019100104.
- [9] N. Gayathri and K. Mahesh, "Improved Fuzzy-Based SVM Classification System Using Feature Extraction for Video Indexing and Retrieval," *Int. J. Fuzzy Syst.*, vol. 22, no. 5, pp. 1716–1729, Jul. 2020, doi: 10.1007/S40815-020-00884-Z.
- [10] R. Jain, P. Jain, T. Kumar, and G. Dhiman, "Real time video summarizing using image semantic segmentation for CBVR," *J. Real-Time Image Process.*, vol. 18, no. 5, pp. 1827–1836, Oct. 2021, doi: 10.1007/S11554-021-01151-6.
- [11] L. Rossetto, I. Giangreco, C. Tănase, and H. Schuldt, "Multimodal video retrieval with the 2017 IMOTION system," *ICMR 2017 - Proc. 2017 ACM Int. Conf. Multimed. Retr.*, pp. 457–460, Jun. 2017, doi: 10.1145/3078971.3079012.
- [12] V. Gabeur, C. Sun, K. Alahari, and C. Schmid, "Multi-modal Transformer for Video Retrieval," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12349 LNCS, pp. 214–229, 2020, doi: 10.1007/978-3-030-58548-8_13.
- [13] X. Cheng, H. Lin, X. Wu, F. Yang, and D. Shen, "Improving Video-Text Retrieval by Multi-Stream Corpus Alignment and Dual Softmax Loss," Sep. 2021, Accessed: Sep. 11, 2024. [Online]. Available: <http://arxiv.org/abs/2109.04290>
- [14] T. H. Chiang, Y. C. Tseng, and Y. C. Tseng, "A multi-embedding neural model for incident video retrieval," *Pattern Recognit.*, vol. 130, Oct. 2022, doi: 10.1016/j.patcog.2022.108807.
- [15] J. A. Portillo-Quintero, J. C. Ortiz-Bayliss, and H. Terashima-Marín, "A Straightforward Framework for Video Retrieval Using CLIP,"

Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12725 LNCS, pp. 3–12, 2021, doi: 10.1007/978-3-030-77004-4_1.

- [16] T. N. Nguyen, L. M. Quang, G. Healy, B. T. Nguyen, and C. Gurrin, "VideoCLIP 2.0: An Interactive CLIP-Based Video Retrieval System for Novice Users at VBS2024," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 14557 LNCS, pp. 394–399, 2024, doi: 10.1007/978-3-031-53302-0_37.
- [17] D. L. Pham *et al.*, "A Deep Learning-Based Real-Time Video Object Contextualizing and Archiving System," *Int. Conf. Adv. Commun. Technol. ICACT*, vol. 2023-Febru, pp. 137–144, 2023, doi: 10.23919/ICACT56868.2023.10079454.
- [18] J. H. Su, Y. T. Huang, H. H. Yeh, and V. S. Tseng, "Effective content-based video retrieval using pattern-indexing and matching techniques," *Expert Syst. Appl.*, vol. 37, no. 7, pp. 5068–5085, Jul. 2010, doi: 10.1016/J.ESWA.2009.12.003.
- [19] J. Dong *et al.*, "Dual Encoding for Video Retrieval by Text," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 8, pp. 4065–4080, Aug. 2022, doi: 10.1109/TPAMI.2021.3059295.
- [20] S. Zhao, L. Zhu, X. Wang, and Y. Yang, "CenterCLIP: Token Clustering for Efficient Text-Video Retrieval," *SIGIR 2022 - Proc. 45th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, pp. 970–981, Jul. 2022, doi: 10.1145/3477495.3531950/SUPPL_FILE/SIGIR22-MODFP0183.MP4.



Dinh-Lam Pham is a research professor at the Contents Convergence Software Research Institute at Kyonggi University, South Korea. He is the vice-chair of Vietnam Innovation Network in Korea, as well. He received B.S. degree and M.S. degree in computer science from Thainguyen University in 2008 and 2010, respectively, and received Ph.D. degree in Computer Science from Kyonggi University in 2021. His research interests include video retrieving system, deep learning, BPM, process mining, deep-learning-based systems.



Lanh, Le Thanh is a lecture and Vice Dean at the Faculty of Information Technology, Dong Nai Technology University, Bien Hoa City, Vietnam. He received his M.S degrees in Electrical and Electronic engineering at Ho Chi Minh City University of Technology and Education, Ho Chi Minh, Viet Nam in 2015. He received his Ph.D in Electrical engineering and Computer Science at National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan in 2020. His research interests include Optimization System, Electrical Electronic Engineering. The Intelligence of Things, Intelligent Power systems, and Optoelectronics, Photonics and Optical System.



Jeonghyeon Chang received the Ph.D. degree in Police Administration from Kyonggi University, South Korea in 2017. From 2019 to 2020, he also worked as Assistant professor with science of international politics in University of Halla. In 2020, he joined the Contents Convergence Software Research Institute, Kyonggi University, South Korea, where he is currently a Research Professor. His interests include crime prevention through active content, community policing, and data mining



Ahn Sang-eun is a researcher at the Content Convergence Software Institute at Kyonggi University in Korea. She received her Bachelor of Arts and Master of Arts degree from Kyonggi University in 2017 and 2019, respectively. Her research interests include text mining, data mining, text-based criminal prevention system.



Kyunghee Sun received the B.E and M.E. degrees in the department of computer science and statistics from Jeju University, Korea, in 1998 and 2008, respectively. She received the Ph.D. degrees in the department of computer science and engineering from Kyunghee University in 2018. From 2006 to 2014, she was a senior researcher at Jeju Technopark. She is currently a research professor in the Contents Convergence Software Research Institute, Kyonggi University, South Korea. Her current research interests include Bigdata platform, Data Fabric, Deep Learning, IT convergence, network protocol, and network security



Internet of Things collaboration architectures and services.

Kwanghoon Pio Kim is a full professor of computer science department and the founder and supervisor of the collaboration technology research laboratory at Kyonggi University, South Korea. He received B.S. degree in computer science from Kyonggi University in 1984. And he received M.S. degree in computer science from Chungang University in 1986. He also received his M.S. and Ph.D. degrees from the computer science department at University of Colorado Boulder, in 1994 and 1998, respectively. At Kyonggi University, he is Dean of the Computerization and Informatics Institute, and the director of the contents convergence software research center, as well. He has also been on the editorial board of the journal of KSII, and the committee member of the several conferences and workshops. His research interests include groupware, workflow systems, BPM, CSCW, collaboration theory, Grid/P2P distributed systems, process warehousing and mining, workflow-supported social networks discovery and analysis, process-aware information systems, data intensive workflows, and process-aware Internet of Things.



Kyoung-Sook Kim is an adjunct professor of the department of computer science and engineering at Kyonggi University, South Korea. She received her B.S. degree in computer science from Kyonggi University in 1984. She also received her M.S. and Ph.D. degrees in computer engineering from Kyunghee University in 2001 and 2019, respectively. She had worked as a researcher and developer at FAST Systems Inc. and OSI Corp. in South Korea. Her research interests include large-scale database systems and applications, enterprise information systems, real-time databases, distributed databases, temporal databases, federate databases, object-oriented database management systems, and the