Journal of information and
communication convergence engineering
http://jicce.org

J. Inf. Commun. Converg. Eng. 20(4): 288-294, Dec. 2022 Regular paper

Automatic Generation of Video Metadata for the Super-personalized Recommendation of Media

Sung Jung Yong, Hyo Gyeong Park, Yeon Hwi You, and Il-Young Moon*, *Member, KIICE*

Department of Computer Science and Engineering, Korea University of Technology and Education, Cheonan 31253, Korea

Abstract

The media content market has been growing, as various types of content are being mass-produced owing to the recent proliferation of the Internet and digital media. In addition, platforms that provide personalized services for content consumption are emerging and competing with each other to recommend personalized content. Existing platforms use a method in which a user directly inputs video metadata. Consequently, significant amounts of time and cost are consumed in processing large amounts of data. In this study, keyframes and audio spectra based on the YCbCr color model of a movie trailer were extracted for the automatic generation of metadata. The extracted audio spectra and image keyframes were used as learning data for genre recognition in deep learning. Deep learning was implemented to determine genres among the video metadata, and suggestions for utilization were proposed. A system that can automatically generate metadata established through the results of this study will be helpful for studying recommendation systems for media super-personalization.

Index Terms: AI, Metadata, OTT, Keyframe, YCbCr

I. INTRODUCTION

With the proliferation of the Internet and digital technology, media service platforms are increasingly storing large amounts of media data and providing customized services online. Metadata must be generated to recommend content that suits an individual's taste. The metadata of the generated content are compared with the user information to provide a personalized service. In addition, the media content market is growing, as various types of content have been mass-produced recently owing to increased accessibility. Content is a product whose diversity is constantly increasing, including content produced by TV broadcasting producers and agencies, original content produced by over-the-top (OTT) services, and content posted on social networking services to attract more users.

Users desire content that suits their taste, and competition for the personalization of content recommendations on various platforms has intensified. Netflix applies its own algorithm that combines content-based filtering and collaborative filtering technologies based on user interests and viewing records [1]. YouTube also uses its own recommendation algorithm based on deep neural networks [2].

High-quality metadata are required for an efficient recommendation system. This is because ultrapersonalized customized services can be provided by matching individual data based on high-quality metadata. Existing platforms use a method in which a user directly inputs image metadata. Consequently, significant amounts of time and cost are consumed in processing large amounts of data.

The content consumed on most OTT platforms is films. Thus, we intend to conduct this study based on movies that are consumed frequently. First, to extract the genre metadata of a movie, we intend to generate metadata automatically

Received 14 April 2022, **Revised** 14 June 2022, **Accepted** 07 July 2022
***Corresponding Author** Il-Young Moon (E-mail: iymoon@koreatech.ac.kr, Tel: +82-41-560-1493)
Department of Computer Science and Engineering, Korea University of Technology and Education, Cheonan 31253, Korea

Open Access **https://doi.org/10.56977/jicce.2022.20.4.288** print ISSN: 2234-8255 online ISSN: 2234-8883

using video keyframes and music, which have a close relationship with the movie. The images of a movie can be recalled by simply listening to the music played in the movie. As such, music has the excellent function of expressing the characteristics of the movie and the emotions of the scene [3].

Therefore, for super-personalized recommendation, we analyzed the audio of movie trailers, extracted the keyframes based on the YCbCr color model, and investigated the process of distinguishing the genre of the movie by analyzing the audio and keyframes of the video.

II. SYSTEM MODEL AND METHODS

In this study, we examine the usability of metadata generation using the audio and video keyframes of movies.

First, we present the composition of the overall system and extract keyframes and music from movie trailers.

In addition, we implement this method with artificial intelligence for distinguishing genres by analyzing the keyframe and audio of movie trailers and confirm the results.

A. Proposal of Video Metadata Extraction System

The flowchart and model design of the system for extracting metadata are proposed. We propose a method for extracting the genre of a movie into metadata. Audio and image data are separated from the movie content, and metadata are extracted from each dataset.

Fig. 1 shows the flowchart of the proposed system. To learn the deep learning model, movie trailers are prepared and separated into audio and image data. The separated image data are used to extract a keyframe using the YCbCr color model. Subsequently, a histogram of the keyframe is generated, and the face recognition model recognizes the face of the movie star to generate metadata about the movie star. The voice is removed from the separated audio data, leaving only the background music, and the audio spectrum of each movie trailer is generated via a short-time Fourier transform (STFT).

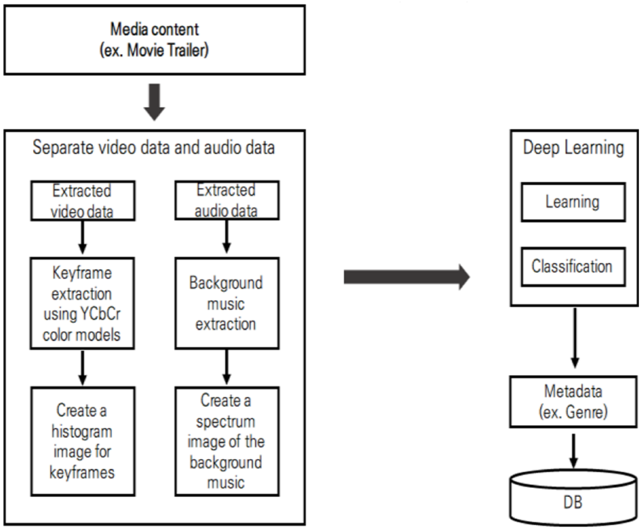

Media content
(ex. Movie Trailer)
Separate video data and audio data
Extracted video data
Extracted audio data
Keyframe extraction using YCbCr color models
Background music extraction
Create a histogram image for keyframes
Create a spectrum image of the background music
Deep Learning
Learning
Classification
Metadata
(ex. Genre)
DB

Fig. 1. Flowchart of the video metadata extraction system

Table 1. Classification of the genres of the movie trailers used

Genre	Trailer 1	Trailer 2	Trailer 3	Trailer 4
Comedy	Extreme Job	Honest Candidate	Ms. Wife	The Secret Zoo
Action	Godzilla: King of the Monsters	Ashfall	Transformers	Iron Mask
Horror	The Nun	Us	Gonjiam: Haunted Asylum	0.0 MHz
Romance	Carol	Notting Hill	Once	The Beauty Inside

Audio spectra and keyframe histograms obtained through a series of processes are stored as images and classified into learning and evaluation data for deep learning. Subsequently, metadata, such as genres, are extracted through deep learning and image classification and stored in a database.

In this study, to overcome the limitations of securing movie data owing to film capacity and copyright problems, we intend to analyze movie trailers and implement the proposed method with artificial intelligence.

As shown in Table 1, trailers were prepared for four films selected for each genre among comedy, action, horror, and romance, and the video and audio data were separated and used.

B. YCbCr Color Model Analysis for Keyframe Extraction

Analysis was performed using Python's matrix or NumPy for multidimensional array processing, OpenCV for image processing, and the PeakUtils library for peak and data detection.

YCbCr is a type of linear color space, where Y represents the luminance component and Cb and Cr represent the concentration offset components of red and blue, respectively. RGB can be converted to YCbCr using the following equation [4]:

$$Y = 0.299R + 0.587G + 0.114B$$
$$CR = 0.212R - 0.523G - 0.311B$$
$$Cb = 0.596R - 0.272G - 0.321B \qquad (1)$$

The process of extracting the keyframe from an image is illustrated in Fig. 2. The RGB color model was converted into a YCbCr color model and the Y, Cb, and Cr color spaces were separated to measure the degree of change. When the degree of change was in the peak state (high state), it was determined that the image was a keyframe image, such as scene conversion, and the keyframe at this time was obtained.

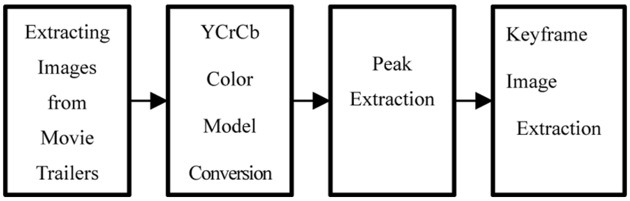
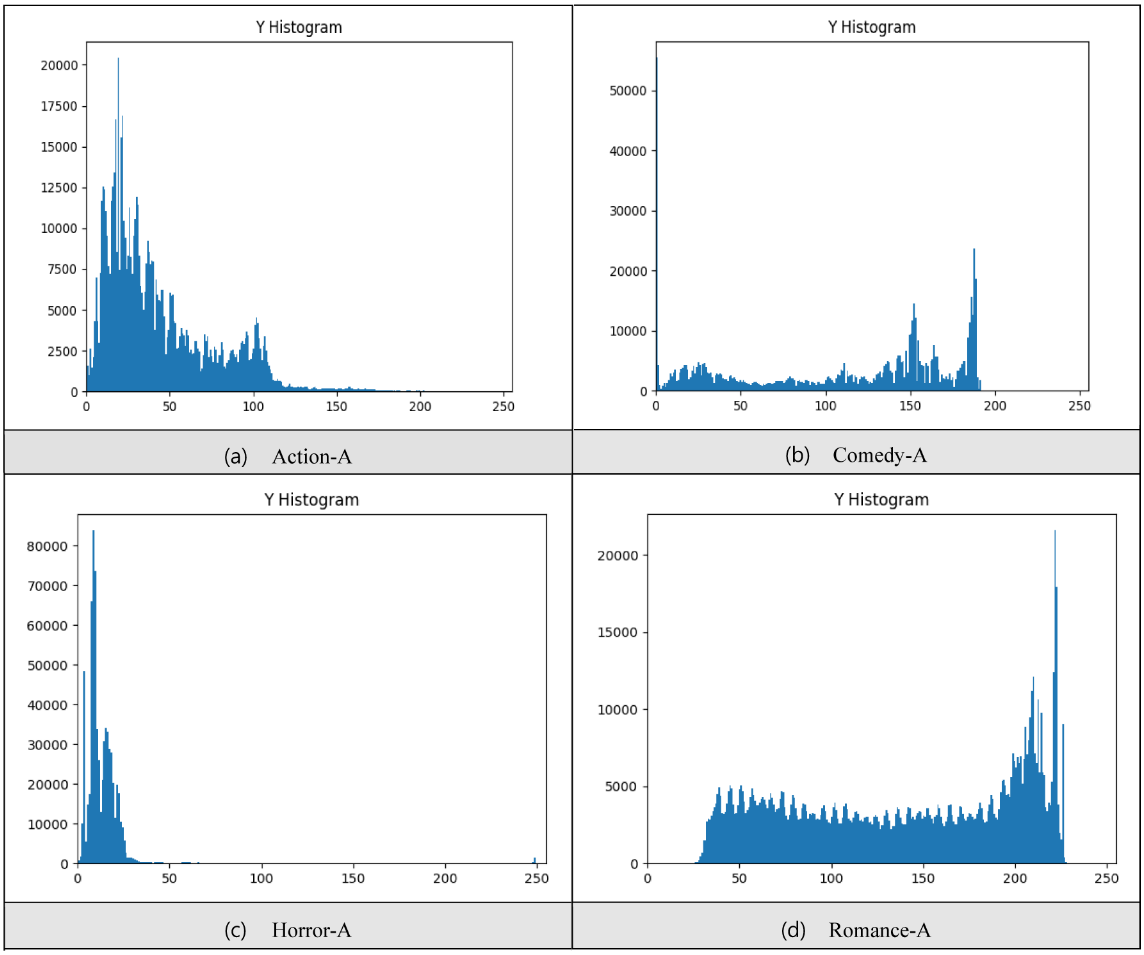

Extracting Images from Movie Trailers
YCrCb Color Model Conversion
Peak Extraction
Keyframe Image Extraction

Fig. 2. Image keyframe extraction process

As shown in Table 2, the genre of the movie was classified into comedy, action, horror, and romance. Four representative movie trailers of each genre were selected, and the video and audio data were separated. Additionally, the RGB color model was converted into a YCbCr color model in the sepa-

Table 2. Movie genre classification and the number of frames extracted

Genre	Movie trailer	Number of keyframes extracted from movie trailer
Action	Action-A	233
	Action-B	215
	Action-C	231
	Action-D	309
Comedy	Comedy-A	118
	Comedy-B	110
	Comedy-C	67
	Comedy-D	63
Horror	Horror-A	50
	Horror-B	152
	Horror-C	125
	Horror-D	92
Romance	Romance-A	172
	Romance-B	177
	Romance-C	111
	Romance-D	53



(a) Action-A

(b) Comedy-A

(c) Horror-A

(d) Romance-A

Fig. 3. Histograms by movie genre

rated image. A keyframe for each movie trailer was generated through the converted YCbCr color space change diagram and stored as an image. Subsequently, 50 keyframes for each movie trailer were extracted randomly. Subsequently, classification by category was performed to process the deep-learning data.

Fig. 3 shows a random selection of histogram images of the Y values extracted using the YCbCr color space. Differences in the histogram may be observed for each movie; however, the histogram derived for each genre had distinct characteristics.

It was expected that, if the histogram image extracted through the YCbCr color model is applied to artificial intelligence, genres can be distinguished based on their characteristics. We attempted to verify the results by implementing this method with artificial intelligence.

C. Artificial Intelligence Application of extracted YCbCr Histogram

As shown in Fig. 4, based on the results of the YCbCr color model, the histograms were applied to convolutional neural networks (CNNs) and logistic regression models to confirm the classification results by genre.

VGG-16 [5], which consists of 16 layers for image classification, was applied, and rectified linear unit (ReLU) was used as the activation function.

D. Audio Analysis of Movie Music

As shown in Fig. 5, audio data extracted from the movie trailers were analyzed to observe the change in the frequency components over time using the STFT [6,7]. The results of this analysis were obtained as spectral images.

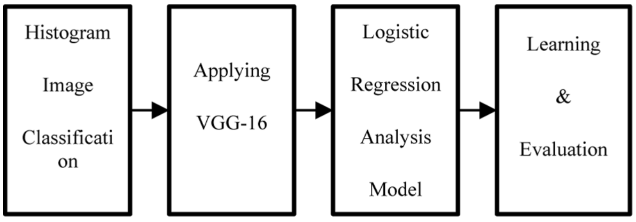

Histogram Image Classificati on
Applying VGG-16
Logistic Regression Analysis Model
Learning & Evaluation

Fig. 4. Implementation of the keyframe analysis with artificial intelligence

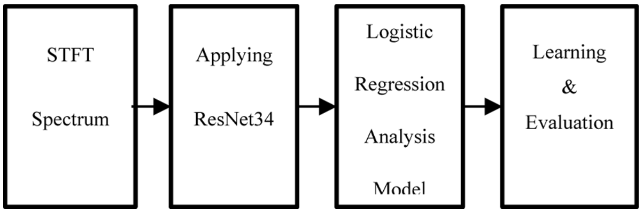


Fig. 5. Audio analysis and artificial intelligence application

The spectral images obtained as a result of audio analysis were classified into learning data by genre, and transfer learning was performed using the ResNet34 [8] deep learning model.

III. RESULTS

A. Results of Artificial Intelligence Application through YCbCr Analysis

As shown in Table 3, it was confirmed that the comedy, action, horror, and romance images were classified using the confusion matrix. In the case of horror and romance, poor classification was observed.

The accuracy and precision were 88.9 and 69.5%, respectively. This result is attributed to the fact that the amount of learning data was relatively small. Accordingly, in future studies, higher accuracy can be achieved if the problem of learning data is addressed. If the keyframes of the movie trailers are extracted to distinguish the genre of the movie, metadata for the genre can be automatically generated.

Table 3. Results of the confusion matrix

	Predicted				
	Comedy	Action	Horror	Romance	Total
Comedy	559	11	66	44	680
Action	50	437	126	57	670
Horror	47	92	448	100	687
Romance	43	72	128	440	683
Total	699	612	768	641	2,720

B. Results of Artificial Intelligence Application through Audio Analysis

For audio signal processing, the STFT was used to analyze the changes in the frequency components over time. The STFT can analyze both time-frequency regions compared with the commonly used fast Fourier transform, resulting in genre-specific background music spectrogram images, as shown in Figs. 6 and 7.

Transfer learning was performed because there was a limitation in the preprocessing of the learning data to acquire spectrograms for movie content and generate an artificial neural network model. Transfer learning was performed using the ResNet34 artificial neural network model, which increased the accuracy to 34 layers by adding a convolution layer to the VGG-19 structure as a skeleton.

Fig. 8 shows the results of the learning data placement for the audio spectrum.

Consequently, it was confirmed that audio spectrum images were well classified according to the genres of

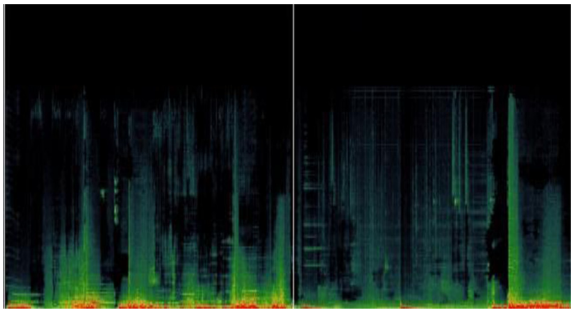


Fig. 6. Spectrogram of background audio for the horror movie trailers

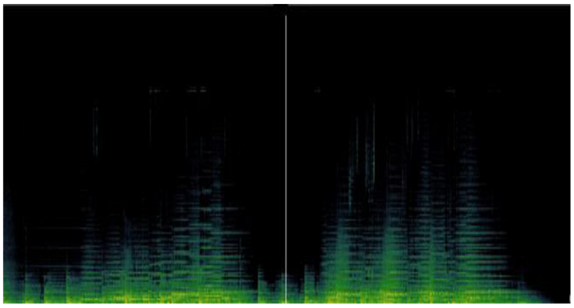


Fig. 7. Spectrogram of background audio for the romance trailers

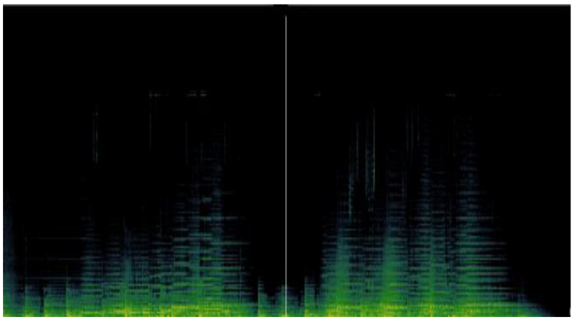


Fig. 8. Visualizing the placement of learning data

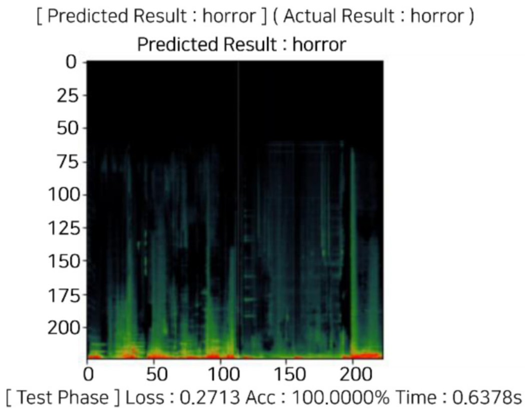


Fig. 9. Testing results

action, horror, romance, and comedy.

Fig. 9 shows the evaluation results of the audio spectrum of the horror genre.

The accuracy was 100%, and the loss value was 0.2713. The high accuracy indicates that the learning data intended to distinguish genres based on the audio spectra were properly recognized and classified. These results confirm that artificial intelligence can distinguish genres using background audio in movies. Table 4 shows the results of evaluation of learning for all the genres.

Thus, metadata for a genre can be automatically generated if the genre is classified.

Table 4. Test data prediction and evaluation results

Training Class		**Test Data**		**Evaluation**		
				Prediction	**Loss**	**Accuracy**
Action	Godzilla	Action	Iron Mask	Action	0.24	100%
	Ashfall					
	Transformers					
Comedy	Extreme Job	Comedy	Secret Zoo	Comedy	0.24	100%
	Ms. Wife					
	Honest Candidate					
Romance	Notting Hill	Romance	Carol	Romance	0.27	100%
	The Beauty Inside					
	Once					
Horror	0.0 MHz	Horror	The Nun	Horror	0.22	100%
	Gonjiam					
	Us					

IV. DISCUSSION AND CONCLUSIONS

Recently, as content has been increasingly mass-produced owing to improved accessibility, the media content market has become more active, and various platforms are actively conducting research on individual metadata and personalized services to satisfy consumers' needs. This paper proposes a method for automatically generating metadata through artificial intelligence, instead of humans directly inputting metadata. First, the keyframes of images were extracted through the YCbCr color model. A histogram image was generated based on the Y value of the extracted keyframe images. There was a difference in the change in the Y value by genre. Then, metadata were automatically generated by implementing the proposed method with CNNs and logistic regression models.

Second, it was confirmed that the STFT spectral images for each genre were extracted through the audio analysis of the movie and applied to the ResNet34 model, resulting in a high accuracy for classification and evaluation after learning. Thus, artificial intelligence was used to generate metadata automatically. Thus, artificial intelligence can automatically generate metadata based on movie elements by using movie keyframes and audio.

In future studies, the type and amount of learning data need to be expanded, and the accuracy of artificial intelligence needs to be improved by extracting the characteristics of keyframe images using not only Y values but also Cr and Cb values. If a system is established to generate metadata automatically through future studies, a recommendation system for media super-personalization can be further developed.

ACKNOWLEDGMENTS

This research was supported by the Basic Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (No. 2021R1I1 A3057800) and this results was supported by "Regional Innovation Strategy (RIS)" through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (MOE) (2021RIS-004).

REFERENCES

[1] C. A. Gomez-Uribe and N. Hunt, "The netflix recommender system: Algorithms, business value, and innovation," *ACM Transactions on Management Information Systems*, vol. 6, no. 13, pp. 1-19, Jan. 2016. DOI: 10.1145/2843948.

[2] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for YouTube recommendations," in *Proceedings of the 10th ACM Conference on Recommender Systems* (RecSys '16), New York: NY, USA, pp. 191-198, 2016. DOI: 10.1145/ 2959100.2959190.

[3] J. Jung, "The correlation of bach music and the scene as seen in films," M. S. thesis, The Graduate School of Ewha Womans University, Seoul, Korea, pp. 1, 2007.

[4] Y. Tan, J. Qin, X. Xiang, W. Ma, W. Pan, and N. Xiong, "A robust watermarking scheme in YCbCr color space based on channel coding," *IEEE Access*, vol. 7, pp. 1-1, Jan. 2019. DOI: 10.1109/ ACCESS.2019.2896304.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint, arXiv: 1409.1556*. Sep. 2014.

[6] Z. Wang, P. Song, Q. Tang, and Y. Rui, "A Non-Stationary Signal Preprocessing Method based on STFT for CW Radio Doppler Signal," in *Proceedings of the 2020 4th International Conference on Vision, Image and Signal Processing*, Bangkok, Thailand, pp. 1-5, 2020. DOI: 10.1145/3448823.3448845.

[7] K. Liu, L. Gong, N. Tian, F. Gong, and Q. Wang, "Feature extraction method of power grid load data based on STFT-CRNN," in *Proceedings of the 6th International Conference on Big Data and Computing (ICBDC'21)*, Shenzhen, Cina, pp. 55-60, 2021. DOI: 10.1145/3469968.3469978.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas: NV, USA, pp. 770-778, 2016. DOI: 10.1109/CVPR.2016.90.

Sung Jung Yong
received a master's degree in computer science and engineering in 2020 from Korea University of Technology and Education, Cheonan, Republic of Korea. He is currently pursuing a Ph.D. from the Department of Computer Science and Engineering at Korea University of Technology and Education. His current research interests are artificial intelligence, web services, and recommendation systems.

Hyo Gyeong Park
received the B.S. degree in computer science and engineering in 2021 from Korea University of Technology and Education, Cheonan, Republic of Korea. She is currently pursuing the M.S. degree from the Department of Computer Science and Engineering at Korea University of Technology and Education. Her current research interests are artificial intelligence, web services, big data, and recommendation systems.

Yeon Hwi You
received the B.S. degree in computer science and engineering in 2022 from Korea University of Technology and Education, Cheonan, Republic of Korea. He is currently pursuing the M.S. degree from the Department of Computer Science and Engineering at Korea University of Technology and Education. His current research interests are artificial intelligence, big data, and recommendation systems

Il-Young Moon
has been a professor at the Department of Computer Science and Engineering, Korea University of Technology and Education, Cheonan, Republic of Korea since 2005. He received the Ph.D. degree from the Department of Aeronautical Communication and Information Engineering, Korea Aerospace University in 2005. His current research interests are artificial intelligence, wireless internet applications, wireless internet, and mobile IP.