Effective Hand Gesture Recognition by Key Frame Selection and 3D Neural Network

Nguyen Ngoc Hoang*, Guee-Sang Lee**, Soo-Hyung Kim**, Hyung-Jeong Yang**

Abstract

This paper presents an approach for dynamic hand gesture recognition by using algorithm based on 3D Convolutional Neural Network (3D_CNN), which is later extended to 3D Residual Networks (3D_ResNet), and the neural network based key frame selection. Typically, 3D deep neural network is used to classify gestures from the input of image frames, randomly sampled from a video data. In this work, to improve the classification performance, we employ key frames which represent the overall video, as the input of the classification network. The key frames are extracted by SegNet instead of conventional clustering algorithms for video summarization (VSUMM) which require heavy computation. By using a deep neural network, key frame selection can be performed in a real-time system. Experiments are conducted using 3D convolutional kernels such as 3D_CNN, Inflated 3D_CNN (I3D) and 3D_ResNet for gesture classification. Our algorithm achieved up to 97.8% of classification accuracy on the Cambridge gesture dataset. The experimental results show that the proposed approach is efficient and outperforms existing methods.

Keywords : hand gesture recognition | dynamic hand gesture | key frame extraction | action recognition

I. INTRODUCTION

Hand gesture recognition is one of the most intuitive and natural ways for communication between human and computational devices. Recently, the role of hand gesture recognition has become more significant in many issues such as human-computer interaction, human-robust interaction and VR/AR applications, due to its convenience and naturalness.

Thanks to the development of computer vision techniques and artificial intelligence, many hand gesture recognition approaches have been developed. In the early years, hand gesture recognition has been handled with conventional machine learning algorithms with handcrafted features [1-3], where 3D skeleton-based geometric features are utilized for classification by a linear classifier SVM [1] or multiclass SVMs [3] are used.

Recently, hand gesture recognition based on a deep convolutional network has been very popular. There are several deep learning frameworks for dynamic hand gesture recognition with temporal-spatial features learning. Typical deep neural networks include 3D_CNN [4], 3D_CNN incorporated with recurrent neural network (RNN) [5], and long-term recurrent convolution network (LRCN) [6].

To improve recognition performance, the idea of utilizing key frame selection with CNN frameworks has been employed for efficient hand gesture recognition [7-10].

In this paper, we focus on dynamic hand gesture classification using 3D neural networks such as 3D_CNN [4], Inflated 3D ConvNet (I3D) [11], 3D_CNN based on ResNet architecture (3D_ResNet) [12]. Since the input includes a set of key frames which are extracted by the fast neural network, based frames key frames. I3D and 3D_ResNet are

* This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A3B05049058 & NRF-2017R1A4A1015559) and by Chonnam National University (Grant number: 2018-3293)

*Student Member, **Member, Professor, Dept. of ECE, Chonnam National University

Manuscript : 2019. 07. 25
Revised : 2019.10. 01
Confirmation of Publication : 2019. 10. 10
Corresponding Author : Guee-Sang Lee,
e-mail : gslee@jnu.ac.kr

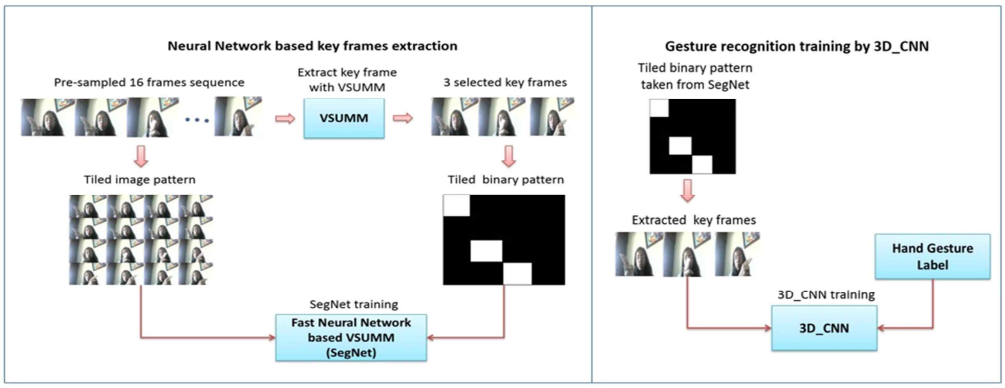


Fig. 1. Overall training phase of our proposed algorithm

the effectiveness of Inception Network [13] and ResNets [14] with 3D convolutional kernels, have been utilized in these networks, respectively. In this algorithm, we apply 3D_CNN type frameworks to classify gestures from a series of important key frames. From gesture image sequences, the key frames are often identified by an efficient video summarization method (VSUMM) [15] based on the K-means clustering. However, VSUMM is expensive in computation time, so it is not suitable for the real-time system. So, the fast neural network is utilized to learn video summarization dictionary that is reasonable for real-time implementation. The deep neural network SegNet is used to estimate key frames and the result is used for gesture classification networks. In our work, the optimal number of key frames per a video sequence is also estimated to get the best classification accuracy.

The main contribution of this paper are: (1) the performance of gesture classification by various 3D_CNN frameworks is enhanced through the incorporation with key framesextraction; (2) the deep neural network SegNet based VSUMM significantly improves the effectiveness of key frame extraction algorithms and appropriate with real-time system; (3) our proposed approach outperforms existing methods on the Cambridge Hand Gesture Dataset; (4) the optimal number of key frames per video to generate best classification performance has been obtained from experiments.

The remainder of this paper is organized as follows. In Section 2, we make a review of relative works. The proposed hand gesture recognition algorithm is presented in section 3. Section 4, experiments conducted on the Cambridge gesture dataset and the SHG dataset are reported. After that, conclusion is given in section 5.

A preliminary version of this paper has been presented in [16]. In this paper, we significantly improved the recognition performance of the algorithm by using 3D_ResNet for classification.

II. PROPOSED METHOD

In our approach, the key frame extraction algorithm selects the important frames of video input, and 3D deep neural network classifies gesture labels from the selected frames. In the training phase, 16 sampled frames of a video data are input to VSUMM to select k key frames. A tiled image pattern is created from 16 sampled frames and a tiled binary pattern is created from selected k key frames as shown in Fig. 1. From the generated tiled image pattern and tiled binary pattern, SegNet [17] learns to identify key frames in a sampled frame sequence. Then classification algorithm based on 3D deep neural network is trained to classify gestures fromthe k key frame input. For the testing phase, from 16 sampled frame input, we create a tiled image pattern and put it into the SegNet based key frame selection. The output of SegNet is tiled binary pattern, from which key frames are extracted. From extracted key frames, the gesture label is identified by 3D neural network as shown in Fig. 2.

2.1 Overall Architecture for Training and Testing

The input video sequence is randomly sampled with 16 frames. The k key frames extracted by VSUMM algorithms are used to create tiled binary patterns for the output label of SegNet. In our algorithm, the number of key frames is set to three (k =3), so the tiled binary pattern contains 16 blocks with 3 non-zero blocks each of which indicates the key frame index. The 3 key frames retrieved from tiled binary patterns of SegNet output are the input to the training network for gesture classification by 3D neural network.

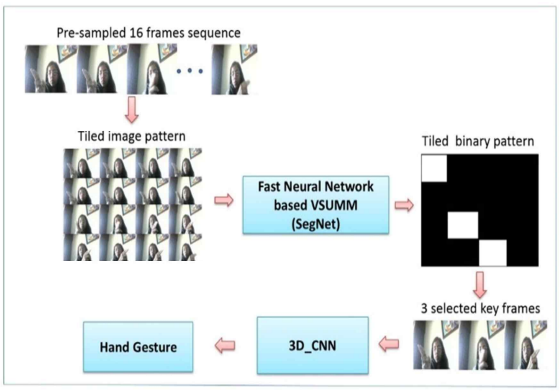


Fig. 2. The overall testing phase of our proposed algorithm

For testing, firstly, 16 frames are randomly sampled from given gesture video input. The sampled frames are resized to generate tiled image pattern with 4 x 4 single blocks. The tiled image pattern is given to SegNet based VSUMM to get the tiled binary pattern which indicate the selected key frames. The selected three keyframes from 3 non-zero blocks are used as input to 3D neural network for the classification of the hand gesture label. The overall procedure of our proposed algorithm for the training and testing phase are shown in Fig. 1 and Fig. 2, respectively.

2.2 Key Frame Extraction by VSUMM

In this approach, we use static video summarization algorithm VSUMM based on K-means clustering as an effective technique for key frame extraction [15]. VSUMM selects *k* key frames from multiple sample frames of the video sequence. From color features of multiple sampled frames, they are clustered into *k* groups, and then the k key frames are selected from each group, as shown in Fig. 3.

In VSUMM, a color histogram is computed in

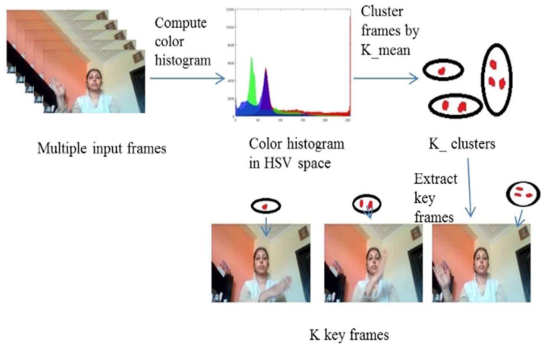


Fig. 3. The key frames extraction by VSUMM

HSV space to normalize the viewing distance of two images by extracting hue histogram. The K-means clustering is used to group similar frames through the distance between frames according to Euclidean distance. For each group, the frame which has the closest distance to the centroid of each cluster is chosen as a key frame. The *k* key frames of a video are selected from each cluster. Because of computation time,

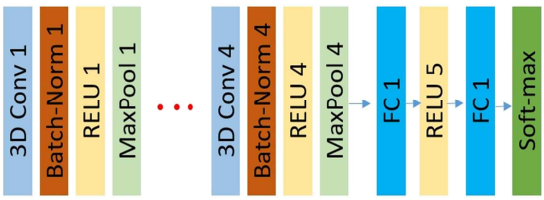


Fig. 4. Basic 3D_CNN flow

the VSUMM algorithm is not appropriate for a real-time system. Instead, SegNet, a deep neural network for fast dictionary learning, is used to approximate the VSUMM algorithm as shown in the next section.

3.3 Training SegNet for Key Frame Selection

The SegNet [17] which is the state-of-art architecture for sematic segmentation is used for key frame extraction by learning dictionary of tiled image patterns and tiled binary patterns. The tiled image pattern is constructed from multiple sampled frames of a video sequence. The tiled binary pattern is a form of multiple non-overlap binary blocks containing indices of k key frames. For training SegNet, the tiled binary pattern created from *k* key frame extracted by VSUMM method is presented as the ground truth.

Because SegNet is one of the fastest frameworks in semantic segmentation, by utilizing it we can accurately approximate key frame selection.

4. 3D Neural Network based classification

For the hand gesture classification, we utilize the 3D neural network to classify gesture labels from images sequences. While 2D_CNNs are known as a framework only suitable for spatial feature learning, 3D_CNNs are known as an effective method for spatiotemporal feature learning. For input as multiple frames, the output of a 3D convolution is multiple frames, while the output of a 2D convolution is an image. With the 2D_CNN, the temporal information

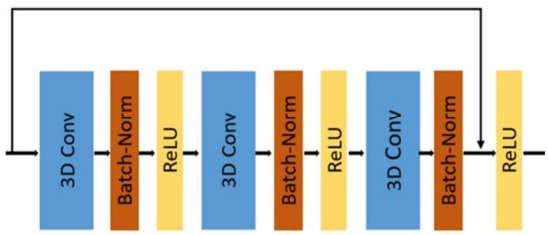


Fig. 5. Example residual block of 3D_ResNet network

will be lost after every convolutional layer, so 3D_CNNs are used for learning by preserving the temporal information in gesture video.

In our approach, the dynamic hand gesture video is first classified by a 3D_CNN, and improved versions of I3D and 3D_ResNet are tried. The basic 3D_CNN used in this approach includes 4 convolutional CONV blocks, 2 fully connected FC layers, a Relu layer, and a soft-max loss layer for label classification, as shown in Fig. 4.

The CONV block is composed of many layers that are 3D convolutional (3D Conv) layer, 3D batch norm, Relu, and max-pooling layer. The number filters in 4 convolutional layers are 64, 128, 256 and 256, respectively. After taking many experiments with various filter size, the 3x3x3 convolution kernel is the best choice for this 3D_CNN algorithm. The 3D Conv layer has kernel size 3x3x3, stride 1 and pad 1 to preserve information of input. The 3D batch norm is added after 3D Conv layer to increase the speed of network by reducing learning rate after several epochs. All max pooling layer have kernel size 2x2x2, except for first max pooling layer. The first pooling layer is max pooling with kernel size 1x2x2 to prevent to merge the temporal feature in the early phase. I3D and 3D_ResNet Network are improved versions of residual 3D_CNN, which are implemented to utilize the effectiveness of Inception Network and ResNets with 3D convolutional kernels, respectively. 3D_ResNet networks consist of many residual convolutional blocks which consist of 3D convolutional (3D Conv) layers, batch norm layers and Relu layer which are structured based on Inception and ResNet architecture. The example of a residual block of the 3D_ResNet network is shown in Fig. 5.

III. EXPERIMENT AND EVALUATION

In this section, experiments are reported for hand gesture recognition by using the 3D_CNN type networks with SegNet for key frame extraction and evaluated on the Cambridge gesture dataset and the SHG dataset.

Table 1. The hand gesture accuracies and computational time of baseline approaches on the SHG dataset

Basic 3D_CNN Algorithm	Average Accuracy (%)	Computation time (ms)
With 18 randomly sampled frames	67.14	5.41
With Manually selected key frames	79.43	4.68
With 3 key frames selected by the proposed algorithm	74.57	26.69
With 4 key frames selected by the proposed algorithm	76.43	27.35
With 5 key frames selected by the proposed algorithm	**77.71**	**28.13**
With 6 key frames selected by the proposed algorithm	74.86	28.48

The Cambridge gesture dataset [18] includes 900 videos about 9 single hand gesture types from three different shape and three motion types in various illuminations. In this dataset, the number of frames per each video around from 36 to 120. The SHG dataset is collected from the 20BN-Jester dataset [20]. The SHG dataset contains 1,400 videos for the training set, and 350 videos for the testing set, that include 7 large-scale hand gesture classes such as: swiping left, swiping right, swiping down, swiping up, zooming in, zooming out, doing other things. Each video of this dataset contains from 35 to 38 RGB frames with variable width. The SHG dataset is more complex than the Cambridge gesture dataset due to the large-scale background and various gestures.

To evaluate the quality of the proposed algorithm, we take experiments various presented 3D_CNN algorithm in some baselines with multiple randomly sampled frames, manually selected 3 key frames, and k key frames selected by only VSUMM method. The number key frames k is considered to estimate the best accuracy classification on the data. The number k changes from 3 to 6.

This experiment is conducted with GPU Geforce GTX 1080 and deep learning framework Keras, python3 library. The gesture recognition accuracy and computation time of those baseline approaches are shown in Table 1.

This proved that our approach has significantly improved the classification accuracy of the basic 3D_CNN algorithm from 67.14% to 77.71%. This

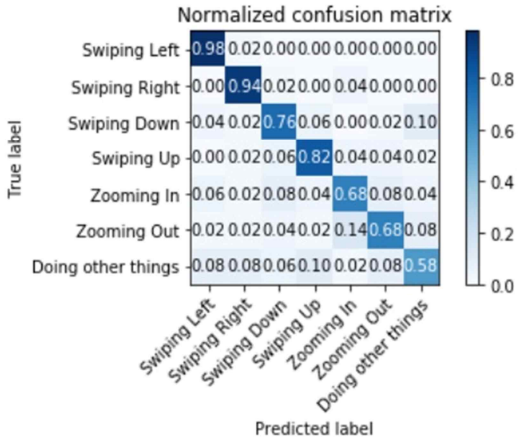


Fig. 6. Confusion matrix for the proposed algorithm on the testing set of the SHG dataset

Table also proves the quality of our key frame extraction algorithm through an approximate comparison with algorithms with manually selected key frames.

In our experiment, the SegNet based key frame extraction takes around 22ms, while key frame extraction by only VSUMM takes 328ms. By using SegNet, the proposed key frame extraction algorithm is clearly faster, and the proposed approach is more suitable for the real-time computational requirement. The average accuracy of the testing experiment conducted on the SHG dataset is 77.71% with 272 correct recognition videos per 350 testing videos.

The confusion matrix for the performance of our proposed algorithm on the testing dataset is shown in Fig. 6.

To evaluate the classification performance, we conducted experiments on the Cambridge gesture dataset and compared results with existing approaches. The comparison is shown in Table 2. The proposed algorithm has the achieved 94.4% accuracy on the Cambridge gesture dataset with 3D_CNN algorithm and it has been improved to 97.8% with 3D_ResNet. Table 2 shows that the proposed approach significantly outperforms existing methods. The basic 3D_CNN based algorithm (94.4% accuracy) has the lower classification accuracy to compare with the improved version I3D (95.7% accuracy) and 3D_ResNet (97.8% accuracy), but it is much faster. By 28.13ms taken time per a video sequence, the experiment proved our system is real-time computational efficiency. Our algorithm is much faster than LSTM_SMRF algorithm proposed at [7] which is 110ms time computation. The experimental results show that our method significantly improved classification accuracy with effectiveness in computation time.

Table 2. The average classification accuracies for the Cambridge gesture dataset

Algorithms	Average classification accuracy (%)	Computation Time (ms)
18 randomly frames_3D CNN	89.5	5.41
16-frame LRCN	86.7	180
Deconv_3D CNN [16]	91.5	78.58
The algorithm proposed at [7]	90.9	110
Our proposed algorithm with 3D_CNN	**94.4**	**28.13**
Our proposed algorithm with I3D	**95.7**	**40.64**
Our proposed algorithm with 3D_ResNet	**97.8**	**41.38**

IV. CONCLUTION

Machine Learning approaches for image processing and computer vision have been studied extensively in recent decades[19,20] and deep learning has been adopted as a major tool for many applications[21,22].

In this paper, we presented a hand gesture recognition method with a combination of 3D neural networks with SegNet based VSUMM. SegNet has been used for efficient key frame extraction and 3D_ResNet is used for classification. The overall performance has been proved to be better than existing methods. For future research, the more robust recognition in the presence of complex backgrounds can be explored.

REFERENCES

[1] Q. D. Smedt; H. Wannous; J.-P. Vandeborr; "Skeleton-Based Dynamic Hand Gesture Recognition", *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2016

[2] U. Cote-Allard; C. L. Fall; A. Campeau-Lecoursy; C. Gosselin; F. Laviolettez; B. Gosselin; "Transfer Learning for sEMG Hand gesture recognition Using Convolutional Neural Networks," *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2017

[3] M. H. Rahman; J. Afrin; "Hand Gesture Recognition using Multiclass Support Vector Machine," *International Journal of Computer Applications*, vol.74, no.1, 2013

[4] D. Tran; L. Bourdev; R. Fergus; L. Torresani; M. Paluri; "Learning spatiotemporal features with 3D convolutional networks," Proc. of IEEE Int. Conf. Comput. Vis. (ICCV), pp.4489-4497, 2015

[5] G. Zhu; L. Zhang; P. Shen; J. Song; "Multimodal Gesture Recognition Using 3-D Convolution and Convolutional LSTM," *IEEE Access*, vol.5, pp.4517-4524, 2017

[6] J. Donahue; L. A. Hendricks; S. Guadarrama; M. Rohrbach; S. Venugopalan; K. Saenko; T. Darrell; "Long-term recurrent convolutional networks for visual recognition and description," *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015

[7] V. John; A. Boyali; S. Mita; M. Imanishi; N. Sanma; "Deep Learning-Based Fast Hand Gesture Recognition Using Representative Frames," *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2016

[8] R. F. Rachmadi; K. Uchimura; G. Koutaki; "Video classification using compacted dataset based on selected keyframe," *IEEE Region 10 Conference (TENCON)*, 2016

[9] H. Tang; H. Liu; W. Xiao; N. Sebe; "Fast and powerful hand gesture recognition extraction and feature fusion," *NeuroComputing* , 2019

[10] H. Jiang; X. Ma; W. Li; S. Ding; C. Mu; "Adaptive key frame extraction from RGB-D for hand gesture recognition," *Tenth International Conference on Digital Image Processing (ICDIP 2018)*, 2018

[11] J. Carreira; A. Zisserman; "Quo Vadis, Action Recognition? A New Model and the Kinetics Datase," *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017

[12] K. Hara; H. Kataoka; Y. Satoh; "Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?," *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018

[13] C. Szegedy; W. Liu; Y. Jiaj; P. Sermanet; S. Reed; D. Anguelov; D. Erhan; V. Vanhoucke; A. Rabinovich; "Going Deeper with Convolutions," *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015

[14] K. He; X. Zhang; S. Ren; J. Sun; "Deep residual learning for image recognition," *Computer Vision and Pattern Recognition (CVPR)*, Proc. of the IEEE Conference on, pp.770-778, 2016

[15] S. E. F. d. Avila; A. P. B. Lopes; A. d. L. Jr.; A. d. A. Arajo; "Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recognition Letters*, vol.32, no.1, pp.56 - 68, 2011

[16] N. N. Hoang; G.-S. Lee; S.-H. Kim; H.-J. Yang; "A Real-time Multimodal Hand Gesture Recognition via 3D Convolutional Neural Network and Key Frame Extraction," *Machine Learning in Medical Imaging (MLMI)*, pp.32-37, 2018

[17] V. Badrinarayanan; A. Kendall; R. Cipolla; "SegNet: A Deep Convolutional Encoder-Decoder Architecture Segmentation," *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014

[18] https://20bn.com/datasets/jester (accessed Mar.,03, 2020).

[19] Abhijeet Boragule; Guee Sang Lee; "Text Line Segmentation of Handwritten Documents by Area Mapping," *Smart Media Journal*, vol.4, no.3, pp.44-49, 2015

[19] Son Tung Trieu; Guee Sang Lee; "Machine Printed and Handwritten Text Discrimination in Korean Document Images," *Smart Media Journal*, vol.5, no.3, pp.30-34, 2016

[20] Tae Seok Lee; Seung Shik Kang; "LSTM based wequence-to-wequence Model for Korean Automatic Word-spacing," *Smart Media Journal*, vol.7, no.4, pp.17-23, 2018

[21] Do Nhu Tai; Soo-Hyung Kim; Guee-Sang Lee; Hyung-Jeong Yang; In-Seop Na; A-Ran Oh; "Tracking by Detection of Multiple Faces using SSD and CNN Features," *Smart Media Journal*, vol.7, no.4, pp.1-69, 2018

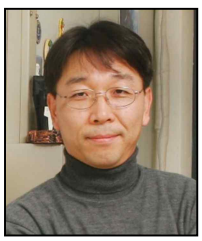
Authors

Hoang-Nam Bui

He received a B.S. degree in Automatic Control from Ha Noi University of Science and Technology, Vietnam in 2017. He is currently a MS student in Electronics and Computer Engineering department of Chonnam National University, Korea. His interests include image processing, video processing, computer vision and pattern recognition.

Guee- Sang Lee

He received a B.S. degree in Electrical Engineering and a M.S. degree in Computer Engineering from Seoul National University, Korea in 1980 and 1982, respectively. He received a Ph.D. degree in Computer Science from Pennsylvania State University in 1991. He is currently a professor of the Department of Electronics and Computer Engineering in Chonnam National University, Korea. His research interests are mainly in the field of image processing, computer vision and video technology.



Soo- Hyung Kim

He received his B.S. degree in Computer Engineering from Seoul National University in 1986, and his M.S. and Ph.D. degrees in Computer Science from Korea Advanced Institute of Science and Technology in 1988 and 1993, respectively. From 1990 to 1996, he was a senior member of research staff in Multimedia Research Center of Samsung Electronics Co., Korea. Since 1997, he has been a professor in the Department of Computer Science, Chonnam National University, Korea. His research interests are pattern recognition, document image processing, medical image processing, and ubiquitous computing.

Hyung- Jeong Yang

She received her B.S., M.S. and Ph.D. degrees from Chonbuk National University, Korea. She was a Post-doc researcher at Carnegie Mellon University, USA. She is currently a professor at Dept. of Electronics and Computer Engineering, Chonnam National University, Gwangju, Korea. Her main research interests include multimedia data mining, pattern recognition, artificial intelligence, e-Learning, and e-Design.