HSFE Network and Fusion Model based Dynamic Hand Gesture Recognition

Do Nhu Tai[1], In Seop Na[2*], Soo Hyung Kim[1]
[1]Department of Computer Science, Chonnam National University, Gwangju, 61186, Korea
[2]SW Convergence Education Institute, Chosun University, Gwangju, 61452, Korea
[e-mail: donhutai@gmail.com, ypencil@hanmail.net, shkim@jnu.ac.kr]
*Corresponding author: In Seop Na

Received April 18, 2019; accepted July 1, 2019; published September 30, 2020

Abstract

Dynamic hand gesture recognition(d-HGR) plays an important role in human-computer interaction(HCI) system. With the growth of hand-pose estimation as well as 3D depth sensors, depth, and the hand-skeleton dataset is proposed to bring much research in depth and 3D hand skeleton approaches. However, it is still a challenging problem due to the low resolution, higher complexity, and self-occlusion. In this paper, we propose a hand-shape feature extraction(HSFE) network to produce robust hand-shapes. We build a hand-shape model, and hand-skeleton based on LSTM to exploit the temporal information from hand-shape and motion changes. Fusion between two models brings the best accuracy in dynamic hand gesture (DHG) dataset.

Keywords: HSFE network, dynamic hand gesture, hand detection, hand gesture recognition, LSTM

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2017R1A4A1015559). And This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF-2018R1A1A1A05022526).

http://doi.org/10.3837/tiis.2020.09.020 **ISSN : 1976-7277**

1. Introduction

Due to the growth of low-cost 3D depth sensors, dynamic hand gesture recognition (d-HGR) has been emerged as an important step in Human-Computer Interaction (HCI) applications, such as sign language recognition, robotics, and interactive gaming. They allow considering 3D information, which leads easily to extract hand region and 3D hand skeletons in the complex environments such as background clutter, occlusions, and light variants [1]. The DHGR is considered typical pattern recognition problems with two steps: feature extraction, and classification.

Up to now, the d-HGR is a challenging task due to its small size, complexity, and self-occlusion. Moreover, it is difficult to recognize because of intra-class dissimilarities, inter-class similarities in gestures. Intra-class gesture dissimilarities come from cultural or individual factors to lead the differences of position, speed, and style of the hand gesture. Inter-class similarities appear in high same among some hand gestures. So, it needs to deal with exploiting the spatial and temporal information of hand gestures to prevent above problems as well as the noise from the device

Traditional handcrafted methods focus on building the robust feature descriptors in the spatial and temporal dimension to encode the changes of hand motion and hand shape such as Histogram of 3D facets [2], Spatio-Temporal HOG2 Descriptor [3], Histogram of Oriented 4D Normals (HON4D) [4], etc.

Besides, with the success of the convolution neural network (CNN) in image classification [5], image segmentation [6] as well as the big dataset ImageNet [7], deep learning also applies in dynamic action/gesture recognition with 2D CNN model [8], 2D CNN integrated motion features [9], 3D CNN model [10], and temporal models such as Long Short-Term Memory (LSTM) [11].

Color and depth stream often are used in previous methods [12]. Some methods use additionally infrared and audio stream, as well as body skeleton [13, 14]. Besides, the rapid development of hand pose estimation [15], it requires the dataset and methods for processing hand skeleton data. De Smedt et al. [16] build Dynamic Hand Gesture Dataset (DHG) for depth and hand skeleton along with the hand-crafted method. From there, many methods based on deep learning [17-21] is proposed the multi-modal from depth and hand skeleton sequences.

HandSegNet[22] and 3D CNN encoder-decoder[23] show that an encoder-decoder architecture is successful in hand pose estimation. Zimmermann et al. [22] use HandSegNet for hand localization, and an encoder-decoder model named PoseNet for estimating key-point score maps. Recently, Moon et al. [23] designs a 3D CNN encoder-decoder to predict 3D pose from a single depth image. From there, the latent codes from encoder-decoder models is the robust features for dynamic hand recognition. However, almost works only focus on the powerful 2D/3D CNN for feature extraction on the specific dataset. Smedt [21] proposes the features from the last layer of 2D CNN hand-pose model for hand-shape representation in dynamic hand-gesture recognition. Input and output of model is depth image and fingertip positions, respectively. Therefore, it is not robust in hand-shape representation to bring efficiently for classification due to the small size, higher complexity and self-occlusions of the hand.

In this paper, we propose the HSFE network to solve the d-HGR problem. We build a hand-shape model and hand-skeleton based on LSTM to exploit the temporal information from hand shape and motion changes. The robust hand-shape feature are extracted by training the hand-shape feature network from the available hand-pose datasets. Our method can handle the complex changes of depth hand sequences by the small size, self-occlusions.
The remainder of the paper consists of four sections. In Section 2, we review the recent related works. In Section 3, we describe our proposed method and its analysis. In Section 4, experiments and discussion are described along with the related methods for comparison. Finally, we conclude our results and discuss further works in Section 5.

2. Related Works

2.1 Overview d-HGR

Hand gesture recognition(HGR) has been rapidly developed in the HCI applications in recent years as follow reasons. Firstly, hand-gestures are intuitive and effective in expressing human feelings. Second, the development of sensor technology has brought hand-gesture such as sensors using accelerometers to capture accurately the movement of the hand and fingertips [25], multi-touch screen sensors widely available through tablets, telephone devices [26], and visual-based sensors [27] for hand-recognition through color images.
The low cost 3D depth sensor such as Microsoft Kinect and Intel RealSense bring many benefits in dealing with HGR more than traditional sensors. Firstly, it is robust to light variants, background clutter, and occlusions. So, it helps easily in hand detection and segmentation. Secondly, the depth sensors capture 3D information in the scene. It helps the development quickly of hand-pose, human-pose estimation in determination the skeleton of human body or hand. Therefore, there are many choices for getting information in HGRs such as depth, color images, and body/hand skeleton [28].
There are two main categories in HGR: static and dynamic HGR. Different from static hand gesture recognition(s-HGR) detecting hand region and extracting hand feature from hand segmentation at the specific time, the dynamic hand gesture recognition(d-HGR) needs to exploit more the temporal features from the hand shape sequences. It treats as the pattern recognition problems consisting of feature extraction, and classification.
Traditional well-known handcrafted features such as HOG, SIFT are extended in depth-base image sequences to describe hand shape feature as well as motion feature. Zhang et al. [2] proposed the Histogram of 3D facets as a depth-based descriptor for s-HGR. Besides, Zhang et al. also use Edge Enhanced Depth Motion Map [29] for encoding shape and motion in d-HGR. Spatio-Temporal HOG2 Descriptor [3] is introduced by Ohn et al. applying in MSR-Hand Gesture Dataset [30]. Oreifej et al. proposed HON4D [4] integrated time, depth and spatial coordinates into 4D space by a histogram of the surface normal orientation distribution. Devane et al. [31] presents skeleton sequences as these trajectories in a Riemannian manifold with action recognition using kNN classification.
A survey of Aghbolaghi et al. [33] depicts the good performance of deep learning approach in action and gesture recognition. The first approach methods [8] uses the power of transfer learning from pre-trained ImageNet [7] of 2D CNN architecture to classify action/gesture recognition by averaging of sampled frames. The second approach methods [9] often use pre-computed motion features for temporal information, while third approach methods [10] take into account 3D convolution as well as 3D pooling. The final approach methods [11]

combines 3D CNN with the temporal sequence model such as Recurrent Neural Network (RNN) and LSTM.

For d-HGR, Molchanov et al. [12] use the volumes of image gradient and depth values for multi-scale 3D-CNN models in VIVA dataset. After that, they improved 3D-CNN models into the Recurrent 3D convolutional neural network [13] with depth, color, and stereo-IR sensors data with successful on ChaLearn Dataset.

2.2 Depth and 3D skeleton d-HGR

With the rapid development of hand pose estimation [15] and supported from depth-based cameras such as Intel RealSense , Microsoft Kinect [34], the hand skeleton features are interesting in the HGR in recent works. Lu et al. [35] use the palm direction, palm normal, fingertips positions and palm center position data from Leap Motion controller to extract features such as fingertip-distances, fingertip-angles, fingertip-elevations, adjacent fingertip-angles for d-HGR. Garcia et al. [14] collected RGB-D sequences as well as hand-pose annotation for first-person hand action recognition. The best base-line method is in merging color, depth and pose data. A multi-modal deep learning framework proposed by Neverova et al. [36] uses color, depth, audio stream as well as body skeleton. The final label of a sequence is computed from voting every frame.
The most recent works, De Smedt et al. [16] published DHG with depth and 2D/3D skeleton information to deal with the lack of benchmark and comparison methods in d-HGR in depth and 3d hand joint approach. They introduce Shape of Connected Joints (SoCJ) descriptor to represent hand shape. After that, Fisher Vector computed from SoCJ descriptor, as well as histograms of the hand direction and the wrist orientation are used for classification. Their method also is the state-of-the-art in handcrafted methods such as HOG2, HON4D, etc.
Moreover, many deep learning methods are also proposed for depth and skeleton d-HGR. Guerry et al. [20] concatenates depth frames randomly to create three key-frames and uses VGG11 [37] as well as pre-trained weights from ImageNet [7] for classification. Chen et al. [17] extracts finger and global motion features from skeleton sequences feeding into a bidirectional recurrent neural network. Devineau et al. [19] builds the parallel convolutions only from hand-skeleton data.

3. Proposed Method

In this section, we describe the proposed method for d-HGR pipeline in **Fig. 1** with normalization hand-depth image, hand-shape feature learning, d-HGR with hand-shape model, and hand-skeleton model, and fusion techniques to combine these models.

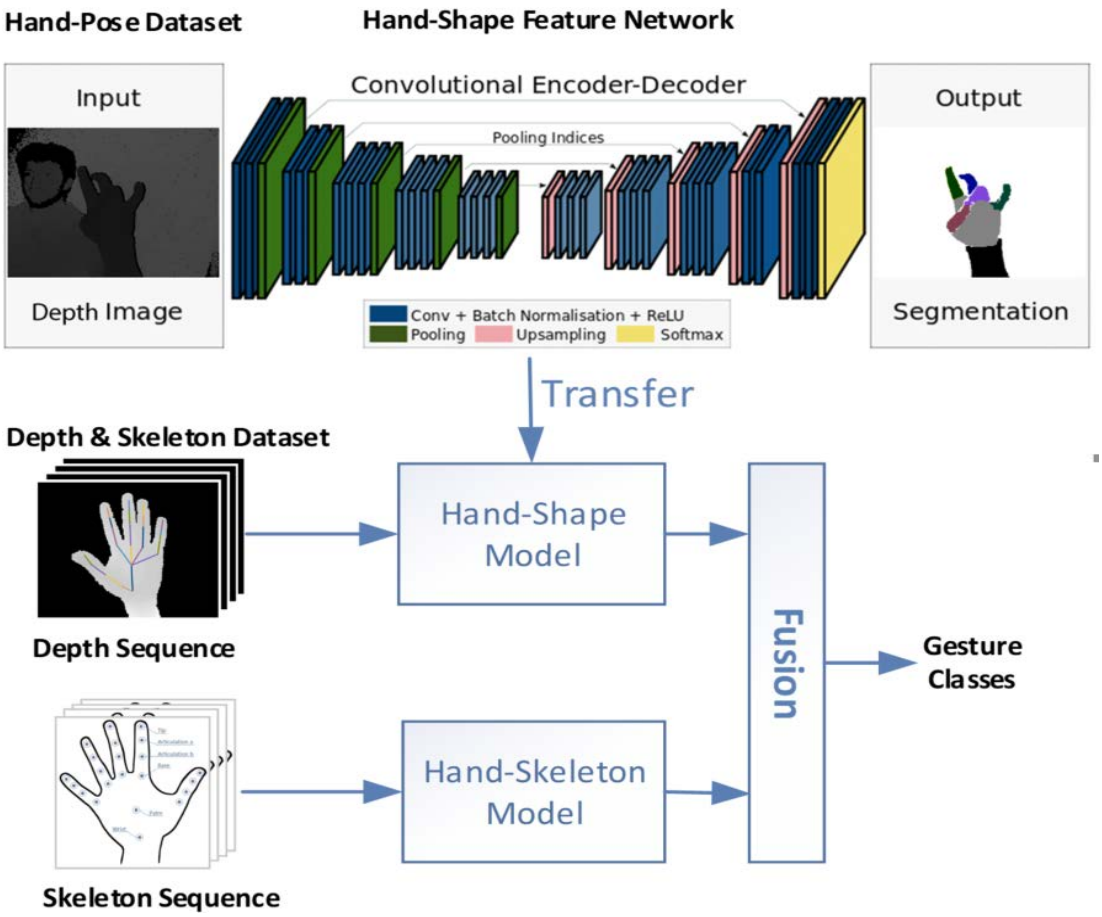

Fig. 1. The pipeline of proposed method

The input of our problem is a hand-depth image sequence $D = \{D_t \in \mathcal{R}^{h \times w}\}_{t=1}^{T}$ and a hand-skeleton sequence$S = \{S_t\}_{t=1}^{T}$, where D_t is the hand-depth image at frame t, $S_t = \{x_i^t, y_i^t, z_i^t\}_{i=1}^{J}$ denotes the hand-skeleton at time t, T is the length of the sequence, and J is the number of hand-skeleton joints. The goal of our problem is to classify $\{D, S\}$ to the gesture c_i with $C = \{c_i\}_{i=1}^{K}$, where C denotes the set of gesture classes in dynamic hand gesture recognition, and K is the amount of classes.

- *We first train hand-shape feature network $\mathcal{H}_{Shape-Feature}(D_t)$ by using SegNet[6] Encoder-Decoder model with FingerPaint dataset[24].*
- *The hand-depth image in the dataset will be normalized before feeding to the network. We use $\mathcal{H}_{Shape-Feature}(D_t)$ to extract hand-shape features of every depth-hand image in depth sequences.*
- *After that, the hand-shape features will be the input of the hand-shape network $\mathcal{H}_{Shape}(D)$ for training and classifying gestures. Due to the arbitrary length of hand depth and skeleton sequences, we will normalize the hand-depth sequence input.*
- *Besides, we also build a hand-skeleton network $\mathcal{H}_{Skeleton}(S)$ receiving the hand-skeleton sequences.*
- *Finally, the result from two model will be integrated by fusion techniques for enhancing the performance the recognition result.*

3.1 Hand-depth image normalization

For a depth image I, we will extract the hand-depth image I_H, where $H = (x, y, w, h)$ is the bounding-box of hand region. We use morphology operator to eliminate the isolated depth pixels. After that, we sort ascending depth values in the hand region and pick depth values in the position range $[d_{min}, d_{max}]$, where d_{min}, d_{max} choose suitable from the experiment. The average value $\mathrm{d_{center}}$ of picked-up depth values will be the center of mass of hand region. With a thresholding t, the depth values in hand region H outside of the range $[d_{center} - t, d_{center} + t]$ are assigned a value of 0. All depth values are normalized to $[0, 255]$.

3.2 HSFE network

HSFE network $\mathcal{H}_{Shape-Feature}(D_t)$ is built from SegNet [6] network with solving image segmentation problem as **Fig. 2**. It is a symmetry network including encoder and decoder parts. The encoder is modified from the VGG16 network with the aim of encoding object into the latent representation. The encoder consists of the blocks of convolutional, batch normalization, ReLU, and Pooling layers. The decoder will responsible for mapping the latent representation of the objects into the semantic tag of the objects. The decoder different from the encoder is to replace Pooling layer by Up-sampling layer. The pooling indices saved from Pooling layers in the encoder is used by corresponding Up-sampling layers to extract location maximum.

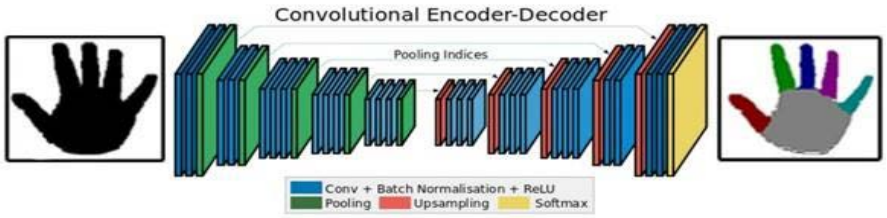

Fig. 2. HSFE network based on SegNet

To extract the robust hand-shape feature, it is carefully to choose the hand-pose dataset for training the hand-shape feature network. Due to the complex attributes of the hand such as small size, self-occlusions, the dataset is suitable to represent the semantic tags of five fingertips, background, and palm region containing the most cases of the hand-poses in the wild.

In this paper, we use FingerPaint dataset [24] as **Fig. 3** for training the hand-shape feature network to transfer pre-train weights into hand-shape network. The FingerPaint dataset consists of captures of five subjects (A, B, C, D, E), with three captures per subject: 'global' for large global movements, relatively static fingers, 'poses' for relatively static global position, moving fingers, and 'combined' for more challenging. It achieves high precision on the pixel-segmentation in pose estimation. In training step, we split 70% every subject for training data and 30% for testing data. We use rotate operator for data augmentation.

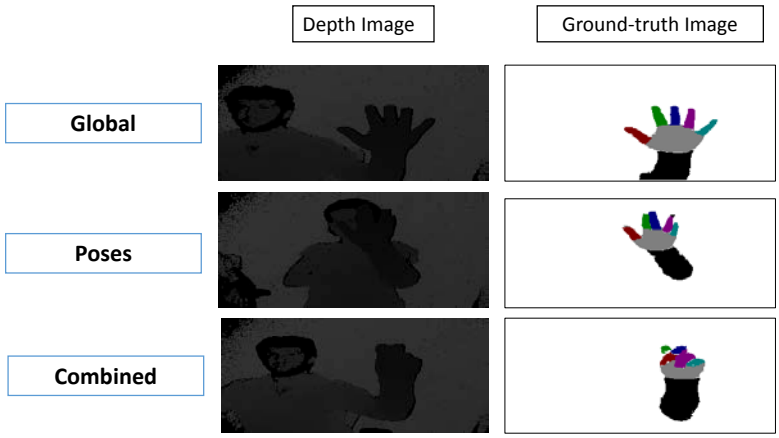

Fig. 3. Example of Finger-Paint dataset[24]

3.3 Hand-Shape and Hand-Skeleton Network

In the hand-shape model, the input sequence is depth hand sequences. Every hand-image region will normalize as mentioned in Section 2.1. After that, we use the last convolution layer in the encoder of hand-shape feature network to extract the hand latent representation. From there, we create the hand-shape feature sequences for the hand-shape network for exploiting the changes in hand-shape by time.

Structure of hand-shape network as **Fig. 4** consists of normalization hand-image, hand-shape feature extraction transferred pre-train weights on Finger-Paint dataset, and the block of LSTM, dense layers and soft-max responsible for exploiting the temporal information of hand pose changes.

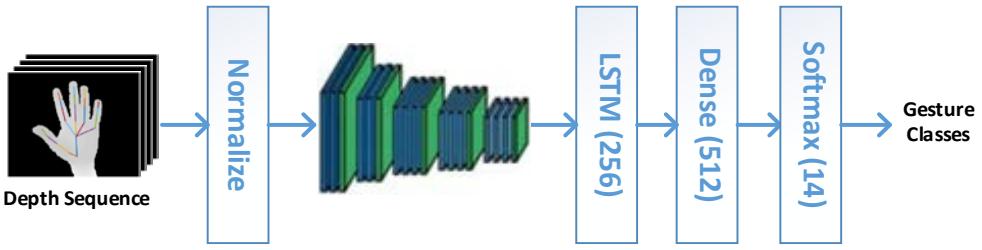

Fig. 4. Hand-Shape Network

Due to the different length of the input sequences, we will normalize them. Firstly, we choose the mean value L of the length of the sequences in DHG dataset as the sequence length of the input data. If the length of the current sequence smaller than L, we use the data at the start and end of sequence for padding. Otherwise, we will choose data randomly satisfying the length is L.

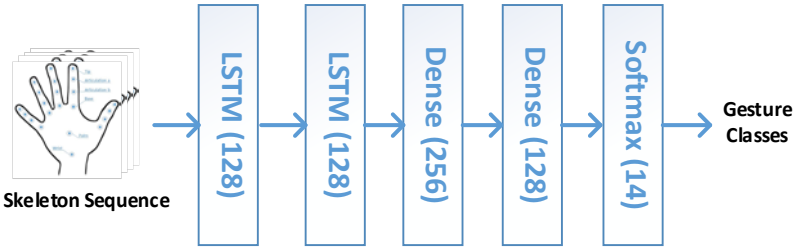

Fig. 5. Hand-Skeleton Network

Skeleton sequences are also normalized as the hand-shape feature sequences as in the hand-shape network. Next, the hand-skeleton will receive them and use the block of two stacked LSTMs, dense layers and soft-max for classification as described in **Fig. 5**.
In two models, we use the Dropout [38] in LSTM as well as in front of soft-max layer. The length for input sequences is 75 based on histogram of the sequence lengths as in **Fig. 6**.

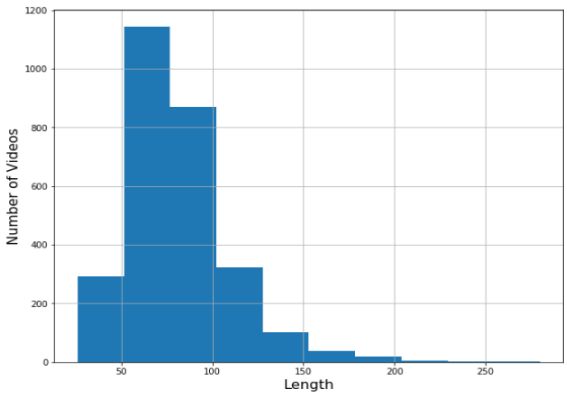

Fig. 6. Histogram of the sequence lengths in DHG Dataset. The mean length is at the value of 75.

3.4 Fusion Techniques

To enhance the performance of the two models, we use fusion techniques, which exploit the complement and redundancy information between the models. In this paper, we use three fusion techniques consisting of late fusion, early fusion and joint fine-tuning as described in [21]. The late fusion technique combines the probability outputs of each deep learning model by majority voting as **Fig. 7**. Given $\hat{y}_{shape}$ and $\hat{y}_{skeleton}$, respectively, the output probabilities of the hand-shape network and hand-skeleton network, the final predicted label are computed as below equation:

$$\hat{y}_{final} = \arg\max_{i}\left(\alpha\hat{y}_{shape} + (1-\alpha)\hat{y}_{skeleton}\right) \quad (1)$$

where $\alpha \in [0,1]$ is the parameter depending on the performance of each network, $i = \overline{1,K}$ with K is the number of gesture classes. In practice, we adjust α from zero to one with step size 0.001 to find the optimal value of the classification result.

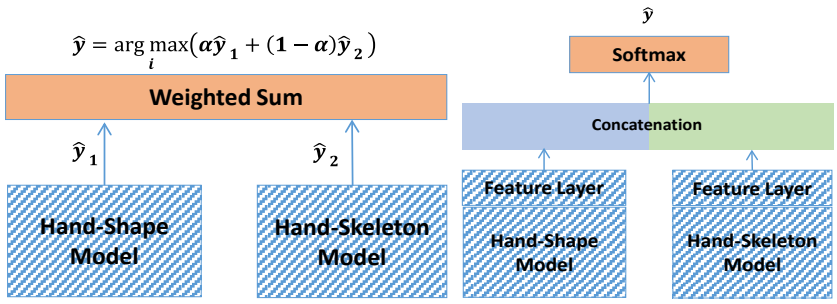

Fig. 7. (Left) weighted sum based fusion technique, and (Right) concatenation based fusion technique

On the other hand, early fusion techniques learn the intermediate feature space from merging the feature space producing from the hand-shape model and hand-shape skeleton as in **Fig. 7**. Finally, joint fine-tuning fusion [39] described in **Fig. 8** integrates two trained models by retraining the last fully connected layers before soft-max with new cost defined as follow:

$$L_{fusion} = \lambda_1 L_{skeleton} + \lambda_2 L_{shape} + \lambda_3 L_{joint} \quad (2)$$

where $L_{skeleton}$, L_{shape} and L_{joint} are loss function computed by hand-skeleton network, hand-shape network and the joint of two networks, respectively. Three parameters λ_1, λ_2, and λ_3 are the control parameters, where λ_1, λ_2 are same, and λ_3 is smaller than λ_1, λ_2.

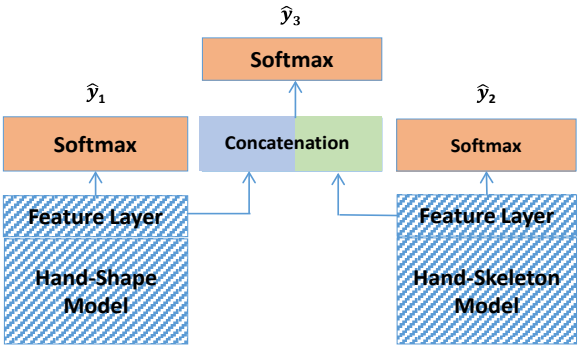

Fig. 8. Joint Fine-Tuning fusion technique using concatenation

$L_{skeleton}$, L_{shape} are the cross entropy loss functions, retrained on linear fully connected network with the last feature layer of the corresponding network connected with soft-max layer. L_{joint} is also the cross entropy loss function of the network constructing from the concatenation two feature layers of two model along with soft-max layer.

4. Experimental and Results

4.1 Environments and Implementation

We built the program on Window environment with Python 3.5. In the program, we used Keras library with Tensorflow backend to develop deep learning models. The experiment was done with the machine's configurations as follows: Intel(R) Core(TM) i7 -8700 CPU @3.20 GHz, GTX 1080Ti 11GB RAM.

We have three main models for training and validating: the HSFE network, hand-shape network, and hand-skeleton network. Besides, we train two fusion models consisting of early and joint fine-tuning fusion. The models is trained first time on Adam algorithm using mini-batch 32, learning rate 0.001 or 0.001. After that, we use Stochastic Gradient Descent (SGD) algorithm with learning rate 0.001 for enhancing performance.

4.2 Datasets and comparison methods

There are two datasets used in our paper including Finger-Paint dataset, and DHG dataset.

On FingerPaint dataset, we use it for HSFE network. The trained weighs is transferred to the hand-shape network for feature extraction. We split on every subject and category of FingerPaint dataset with 70% for training and 30% for validating. DHG dataset with 14 or 28 gesture classes as **Table 1** is applied in d-HGR.

Table 1. Gesture list in DHG dataset

14 classes	28 classes	Gesture	Label
1	1, 2	Grab	Fine
2	3, 4	Tap	Coarse
3	3, 4	Expand	Fine
4	5, 6	Pinch	Fine
5	7, 8	Rotation CW	Fine
6	9, 10	Rotation CCW	Fine
7	13, 14	Swipe Right	Coarse
8	15, 16	Swipe Left	Coarse
9	17, 18	Swipe Up	Coarse
10	19, 20	Swipe Down	Coarse
11	21, 22	Swipe X	Coarse
12	23, 24	Swipe V	Coarse
13	25, 26	Swipe +	Coarse
14	27, 28	Shake	Coarse

DHG dataset has 2800 sequences with 20 participants for 5 trials in 2 ways depending on the number of fingers with one finger and the whole hand. The depth images and hand skeletons were received from Intel RealSense camera. We also split DHG dataset with 70% for training and 30% for validating.

4.3 HSFE

We compare our method with Taylor et al. [40] using a smooth model of the hand, Sharp et al. [24] with the pipeline hand-tracker, and Tan et al. [41] using the 5 different shape models to personalize to each subject. We quantify the error of classification based on counting the percentage of the dataset with the average pixel classification error rate below a specific threshold. It shows the fully and accurately to segment every pixel.

Our result in **Fig. 9** shows our hand-segmentation model better than with Tan et al. method and Sharp et al. method. The strength of our model is to apply in HSFE to use in d-HGR.

We experimented HSF network on FingerPaint dataset with the prediction result shown in **Table 2.** The result of hand segmentation is the good comparison with ground-truth. In the difficult case in row 3 of **Table 2**, it shows exactly for hand-arm and palm region. Some fingers with self-occlusion are not so good.

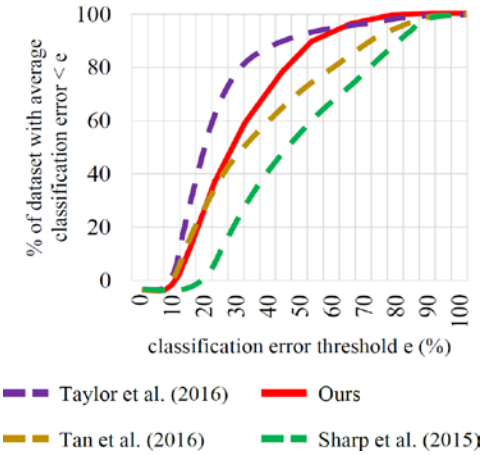

Fig. 9. Classification error on FingerPaint dataset

Table 2. Prediction result on finger-paint dataset

Input Image	Ground-truth	Prediction
		
		
		
		
		

4.4 d-HGR

We evaluated with the hand-shape, hand-skeleton model as well as three fusion techniques as late fusion, early fusion, and joint fine-tuning fusion. The accuracy of every model per gesture class is shown in **Table 3.**

Table 3. The accuracy of the models for classification 14 gesture classes on DHG dataset

Gesture	Shape	Skeleton	Late Fusion	Early Fusion	Joint Fusion
Grab	69	74	69	88	90
Tap	92	70	92	95	95
Expand	95	92	98	95	98
Pinch	82	76	87	87	89
Rotation CW	94	85	94	91	94
Rotation CCW	100	90	100	95	98
Swipe Right	81	86	84	91	86
Swipe Left	95	98	95	98	98
Swipe Up	89	76	89	95	97
Swipe Down	88	95	95	100	98
Swipe X	91	98	94	98	98
Swipe V	72	91	81	97	97
Swipe +	91	94	94	94	97
Shake	93	93	93	100	100
Overall	88.39	87.32	90.54	94.64	**95.36**

In 14 gesture classification, the hand-shape model shows good accuracy with 88.39% better than the hand-skeleton model with 87.32%. The gesture Grab/Pinch has a wide confusion shown in **Fig. 10** because of very similar and only different with the amplitude of the hand movement. The gesture Swipe Right and Swipe V also has high confusion up to 12%. The accuracy of almost gestures is greater than 80%.

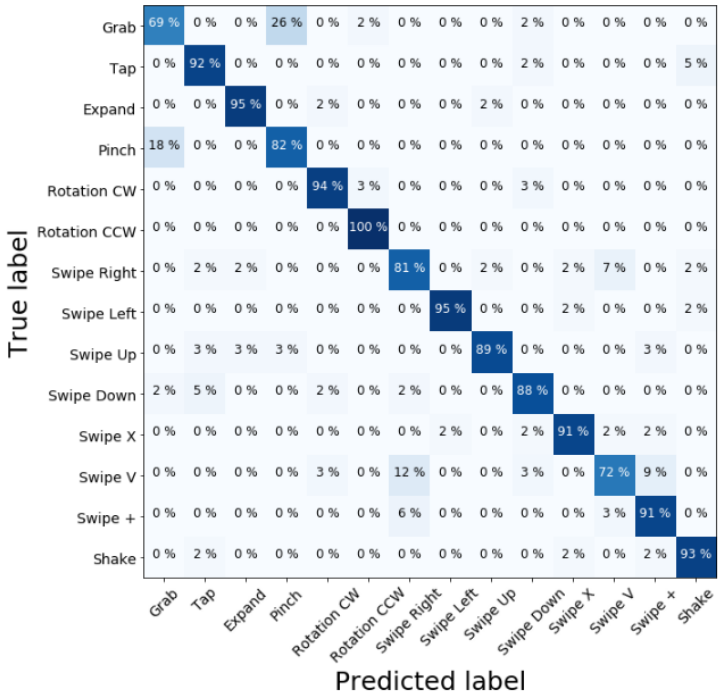

Fig. 10. Confusion matrix of hand-shape model with 14 gesture classes for the accuracy 88.39%

We experimented on DHG dataset with 28 gesture classes with the accuracy shown in **Table 4.** The best accuracy method is at early fusion technique.

Table 4. The accuracy of the models for classification 28 gesture classes on DHG dataset

Gesture	Shape	Skeleton	Late Fusion	Early Fusion	Joint Fusion
Grab (1)	72	64	76	76	80
Grab (2)	71	71	76	76	16
Tap (1)	100	85	100	100	100
Tap (2)	81	81	81	100	100
Expand (1)	92	64	92	96	96
Expand (2)	93	100	93	100	100
Pinch (1)	83	50	78	78	78
Pinch (2)	85	75	85	90	90
Rotation CW (1)	95	85	95	100	95
Rotation CW (2)	86	71	93	93	93
Rotation CCW (1)	100	78	100	89	94
Rotation CCW (2)	100	82	100	100	100
Swipe Right (1)	92	65	92	92	88
Swipe Right (2)	71	59	71	82	65

Swipe Left (1)	95	79	95	100	100
Swipe Left (2)	96	92	96	100	100
Swipe Up (1)	90	65	90	95	95
Swipe Up (2)	94	83	94	94	100
Swipe Down (1)	95	80	95	100	100
Swipe Down (2)	95	91	100	100	100
Swipe X (1)	92	92	92	96	96
Swipe X (2)	78	96	87	96	96
Swipe V (1)	92	83	92	100	100
Swipe V (2)	90	90	90	95	95
Swipe + (1)	88	76	88	94	94
Swipe + (2)	100	89	100	100	100
Shake (1)	88	85	88	92	92
Shake (2)	89	95	95	100	100
Overall	89.29	79.46	90.36	**94.11**	93.75

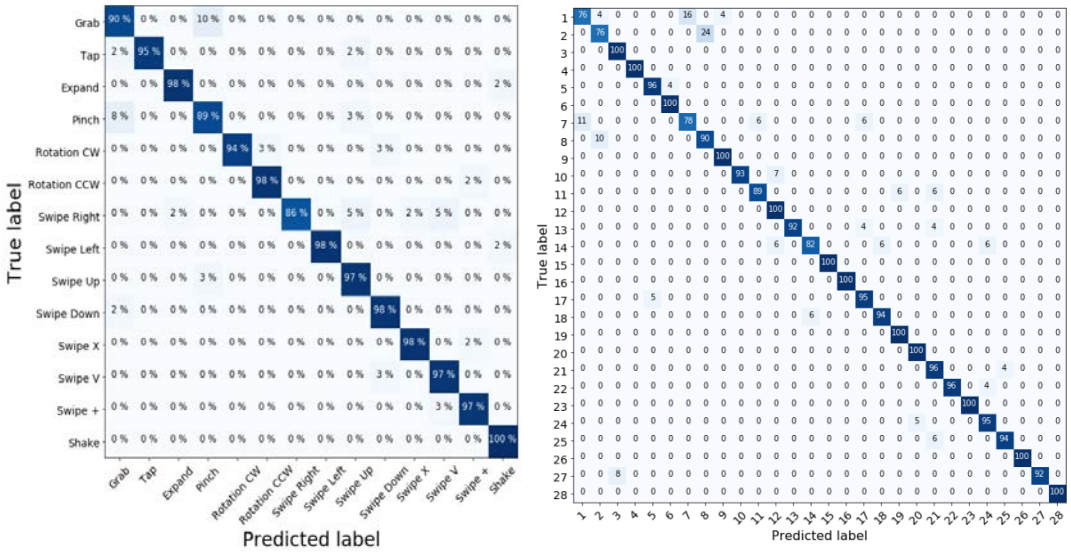

Fig. 11. (a) Confusion matrix of joint fine-tuning model with 14 gesture classes for the accuracy 95.36%, (b) Confusion matrix of early fusion model with 28 gesture classes for the accuracy 94.1%

The hand-shape model at 24 gesture classes also gives the good accuracy with 89.29%. It shows that the hand-shape features are good under hand-shape changes. The accuracy of the hand-skeleton model is 79.46% decreasing 8% due to decreasing the information about the number of fingers on the gestures. From the resulting confusion matrix of the best methods in 14 and 28 gesture classes in **Fig.11.a** and **Fig.11.b** respectively, it shows the combine fusion techniques of two models to give the best accuracy. Early fusion and joint fusion techniques exploit the complement between hand-shape and hand-skeleton feature representation.

Finally, we compare our proposed method with the related methods in traditional methods and deep learning methods. Our proposed method gives the best accuracy more than the remaining methods as in **Table 5.**

Table 5. The Accuracy of Gesture in DHG dataset

Method	14 gestures (%)	28 gestures (%)
Traditional hand-crafted approach		
Ohn-Bar et al. [3]	83.85%	76.53%
Oreifej et al. [4]	78.53%	74.03%
Devanne et al. [31]	79.61%	62.00%
De Smedt et al. [16]	88.24%	81.90%
Deep learning approach		
Guerry et al. [20]	82.90%	71.90%
De Smedt et al. [21]	94.17%	90.48%
Chen et al. [17]	84.68%	80.32%
Núñez et al. [18]	85.60%	81.10%
Our proposed method	**95.36%**	**94.11%**

5. Conclusion

In this paper, we propose a d-HGR method with depth and skeleton classification approach based on the HSFE network and fusion between hand-shape model and hand-skeleton model.

Firstly, the system trained the hand-shape feature model on FingerPaint dataset to extract features of every depth-hand image. Afterward, the hand-shape model exploited temporal information from the hand-shape changes of depth sequence. Our experimental results corroborate that the hand-shape features can cope with various complexity, low-resolution, and self-occlusion of hand-shape changes in the gestures. It takes the accuracy 88.39% for classification 14 gesture classes, and 89.29% for classification 28 gesture classes. It shows that the hand-shape model gives good accuracy when increasing the number of gesture classes.

Besides, the system built the hand-skeleton model to exploit temporal information of hand-pose changes. The accuracy of model is 87.32 for classification 14 gesture classes, and 79.46% for classification 28 gesture classes. The reason for decreasing 8% accuracy is due to decreasing the information about the number of fingers on the gestures.

To boost up the accuracy of the overall system, hand-shape and hand-skeleton models are integrated by fusion techniques such as weighted sum, early fusion, and joint fine-tuning fusion. The accuracy of the model is 90.54% (90.36%), 94.64% (94.11%), and 95.36% (93.75%) corresponding with the weighted sum, early fusion, and joint fine-tuning fusion for classification 14 (28) gesture classes. With the above result, our proposed method achieves the best accuracy on DHG dataset comparing with the traditionally handcrafted methods well as deep learning methods.

In future works, we need to improve the hand-skeleton model to enhance accuracy when decreasing the information about the number of fingers on the gestures.

References

[1] F. Coleca, T. Martinetz, and E. Barth, “Gesture interfaces with depth sensors,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8200 LNCS, pp. 207–227, 2013. Article (CrossRef Link)

[2] C. Zhang and Y. Tian, “Histogram of 3D Facets: A depth descriptor for human action and hand gesture recognition,” *Computer Vision and Image Understanding.*, vol. 139, pp. 29–39, 2015. Article (CrossRef Link)

[3] E. Ohn-Bar and M. M. Trivedi, “Joint Angles Similarities and HOG2 for Action Recognition,” in *Proc. of the IEEE conference on computer vision and pattern recognition workshops*, pp. 465–470, 2013. Article (CrossRef Link)

[4] O. Oreifej and Z. Liu, "HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences," in *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 716–723, 2013. Article (CrossRef Link)

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017. Article (CrossRef Link)

[6] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017. Article (CrossRef Link)

[7] O. Russakovsky, J.Deng, H.Su, J.Krause, S.Satheesh, S.Ma, Z.Huang, and A.Karpathy, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015. Article (CrossRef Link)

[8] L. Sun, K. Jia, D. Yeung, and B. E. Shi, "Human Action Recognition Using Factorized Spatio-Temporal Convolutional Networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4597–4605, 2015. Article (CrossRef Link)

[9] G. Varol, I. Laptev, and C. Schmid, "Long-Term Temporal Convolutions for Action Recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1510–1517, 2018. Article (CrossRef Link)

[10] Z. Liu, C. Zhang, and Y. Tian, "3D-based Deep Convolutional Neural Network for action recognition with depth sequences," *Image and Vision Computing*, vol. 55, pp. 93–100, 2016. Article (CrossRef Link)

[11] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. of the IEEE conference on computer vision and pattern recognition*, vol. 07–12–June, pp. 1110–1118, 2015. Article (CrossRef Link)

[12] P. Molchanov, S. Gupta, K. Kihwan, and J. Kautz, "Hand gesture recognition with 3D convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 1–7, 2015. Article (CrossRef Link)

[13] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz, "Online Detection and Classification of Dynamic Hand Gestures with Recurrent 3D Convolutional Neural Networks," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4207–4215, 2016. Article (CrossRef Link)

[14] G. Garcia-Hernando, S. Yuan, S. Baek, and T.-K. Kim, "First-Person Hand Action Benchmark with RGB-D Videos and 3D Hand Pose Annotations," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 409–419, 2017. Article (CrossRef Link)

[15] J. S. Supancic, G. Rogez, Y. Yang, J. Shotton, and D. Ramanan, "Depth-Based Hand Pose Estimation: Data, Methods, and Challenges," in *Proc. of the IEEE international conference on computer vision*, pp.1868-1876, 2015. Article (CrossRef Link)

[16] Q. De Smedt, H. Wannous, and J. P. Vandeborre, "Skeleton-Based Dynamic Hand Gesture Recognition," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1206–1214, 2016. Article (CrossRef Link)

[17] X. Chen, H. Guo, G. Wang, and L. Zhang, "Motion feature augmented recurrent neural network for skeleton-based dynamic hand gesture recognition," in *Proc. of IEEE International Conference on Image Processing (ICIP)*, pp. 2881–2885, 2017. Article (CrossRef Link)

[18] J. C. Núñez, R. Cabido, J. J. Pantrigo, A. S. Montemayor, and J. F. Vélez, "Convolutional Neural Networks and Long Short-Term Memory for skeleton-based human activity and hand gesture recognition," *Pattern Recognition*, vol. 76, pp. 80–94, 2018. Article (CrossRef Link)

[19] G. Devineau, F. Moutarde, W. Xi, and J. Yang, "Deep learning for hand gesture recognition on skeletal data," in *Proc. of IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pp. 106–113, 2018. Article (CrossRef Link)

[20] Q. De Smedt, H. Wannous, J. Vandeborre, J. Guerry, B. Le Saux, and D. Filliat, "SHREC'17 Track : 3D Hand Gesture Recognition Using a Depth and Skeletal Dataset," *3DOR - 10th Eurographics Workshop on 3D Object Retrieval*, pp. 1-6, 2017. Article (CrossRef Link)

[21] Q. De Smedt, "Dynamic hand gesture recognition - From traditional handcrafted to recent deep learning approaches," *Computer Vision and Pattern Recognition* [cs.CV], Université de Lille 1, Sciences et Technologies; CRIStAL UMR 9189, 2017.

[22] C. Zimmermann and T. Brox, "Learning to Estimate 3D Hand Pose from Single RGB Images," in *Proc. of the IEEE International Conference on Computer Vision*, pp. 4903–4911, 2017. Article (CrossRef Link)

[23] J. Y. Chang, G. Moon, and K. M. Lee, "V2V-PoseNet: Voxel-to-Voxel Prediction Network for Accurate 3D Hand and Human Pose Estimation from a Single Depth Map," in *Proc. of*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5079–5088, 2018. Article (CrossRef Link)

[24] T. Sharp et al., "Accurate, Robust, and Flexible Real-time Hand Tracking," in *Proc. of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 3633–3642, 2015. Article (CrossRef Link)

[25] X. Zhang, X. Chen, Y. Li, V. Lantz, K. Wang, and J. Yang, "A framework for hand gesture recognition based on accelerometer and EMG sensors," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 41, no. 6, pp. 1064–1076, 2011. Article (CrossRef Link)

[26] H. Olafsdottir and C. Appert, "Multi-touch gestures for discrete and continuous control," in *Proc. of the 2014 International Working Conference on Advanced Visual Interfaces*, pp.177-184, 2014. Article (CrossRef Link)

[27] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey," *Artificial Intelligence Review*, vol. 43, no. 1, pp. 1–54, 2015. Article (CrossRef Link)

[28] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A Survey," *Pattern Recognition Letters*, vol. 119, pp. 3–11, 2019. Article (CrossRef Link)

[29] C. Zhang and Y. Tian, "Edge enhanced depth motion map for dynamic hand gesture recognition," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 500–505, 2013. Article (CrossRef Link)

[30] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu, "Robust 3D action recognition with random occupancy patterns," in *Proc. of European Conference on Computer Vision*, pp. 872–885, 2012. Article (CrossRef Link)

[31] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, and A. Del Bimbo, "3-D Human Action Recognition by Shape Analysis of Motion Trajectories on Riemannian Manifold," *IEEE transactions on cybernetics*, vol. 45, no. 7, pp. 1340–1352, 2015. Article (CrossRef Link)

[32] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp.1137-1149, 2017. Article (CrossRef Link)

[33] M. Asadi-Aghbolaghi *et al.*, "A Survey on Deep Learning Based Approaches for Action and Gesture Recognition in Image Sequences," in *Proc. of IEEE international conference on automatic face & gesture recognition (FG)*, pp. 476–483, 2017. Article (CrossRef Link)

[34] L. A. Anonymous, E. Krupka, N. Bloom, D. Freedman, A. Vinnikov, and A. B. Hillel, "Toward realistic hands gesture interface : Keeping it simple for developers and machines," in *Proc. of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 1887–1898, 2017. Article (CrossRef Link)

[35] W. Lu, Z. Tong, and J. Chu, "Dynamic hand gesture recognition with leap motion controller," *IEEE Signal Processing Letters*, vol. 23, no. 9, pp. 1188–1192, 2016. Article (CrossRef Link)

[36] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, "ModDrop: Adaptive Multi-Modal Gesture Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1692–1706, 2016. Article (CrossRef Link)

[37] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. of the IEEE conference on computer vision and pattern recognition*, pp.1-9, 2015. Article (CrossRef Link)

[38] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *The Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[39] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proc. of the IEEE international conference on computer vision*, pp. 2983–2991, 2015. Article (CrossRef Link)

[40] J. Taylor *et al.*, "Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, pp. 1–12, 2016. Article (CrossRef Link)

[41] D. J. Tan *et al.*, "Fits Like a Glove: Rapid and Reliable Hand Shape Personalization," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5610–5619, 2016. Article (CrossRef Link)

Do Nhu Tai received the B.S. in Information System major from HCM City University of Foreign Language – Information Technology, Vietnam in 2005 and his M.S. in Information System Management from International University, Vietnam National University at HCMC, Viet Nam in 2017. From 2005 to 2017, he was a lecturer in Faculty of Information Technology, HCM University of Foreign Languages and Information Technology, Vietnam. Since 2017, he is Ph.D candidate in the Department of Computer Science, Chonnam National University, Korea. His research interests are pattern recognition, deep learning, computer vision, and parallel programming.

In Seop Na received his B.S., M.S. and Ph.D. degrees in Computer Science from Chonnam National University, Korea in 1997, 1999 and 2008, respectively. From 2012 to 2018, he was a research professor in Chonnam National University, Korea. From Jan. to May 2018, he was a visiting scholar in University of California (Merced), USA. Since 2018, he has been an assistant professor in Software Convergence Education Institute, Chosun University, Korea. His research interests are visual intelligence, aritificial intelligence, image processing, pattern recognition, object detection/segmentation/recognition and tracking, human emotion recognition.

Soo Hyung Kim 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.