

# Quantitation of new arbitrary view dynamic human action recognition framework

Anh-Dung Ho<sup>1</sup>, Huong-Giang Doan<sup>2</sup>

<sup>1</sup>Department of Information Technology, East Asia University of Technology, Ha Noi, Viet Nam

<sup>2</sup>Faculty of Control and Automation, Electric Power University, Ha Noi, Viet Nam

---

## Article Info

### Article history:

Received Jun 21, 2024

Revised Oct 3, 2024

Accepted Oct 7, 2024

### Keywords:

Arbitrary view recognition

Convolution neural network

Deep learning

Generative adversarial networks

Human activity recognition

Multiple view recognition

## ABSTRACT

Dynamic action recognition has attracted many researchers due to its applications. Nevertheless, it is still a challenging problem because the diversity of camera setups in the training phases are not similar to the testing phases, and/or the arbitrary view actions are captured from multiple viewpoints of cameras. In fact, some recent dynamic gesture approaches focus on multiview action recognition, but they are not resolved in novel viewpoints. In this research, we propose a novel end-to-end framework for dynamic gesture recognition from an unknown viewpoint. It consists of three main components: (i) a synthetic video generation with generative adversarial network (GAN)-based architecture named ArVi-MoCoGAN model; (ii) a feature extractor part which is evaluated and compared by various 3D CNN backbones; and (iii) a channel and spatial attention module. The ArVi-MoCoGAN generates the synthetic videos at multiple fixed viewpoints from a real dynamic gesture at an arbitrary viewpoint. These synthetic videos will be extracted in the next component by various three-dimensional (3D) convolutional neural network (CNN) models. These feature vectors are then processed in the final part to focus on the attention features of dynamic actions. Our proposed framework is compared to the SOTA approaches in accuracy that is extensively discussed and evaluated on four standard dynamic action datasets. The experimental results of our proposed method are higher than the recent solutions, from 0.01% to 9.59% for arbitrary view action recognition.

*This is an open access article under the [CC BY-SA](#) license.*



---

## Corresponding Author:

Huong-Giang Doan

Faculty of Control and Automation, Electric Power University

235 Hoang Quoc Viet, Ha Noi, Viet Nam

Email: giangdth@epu.edu.vn

---

## 1. INTRODUCTION

Human activity recognition (HAR) has become an attractive field in computer vision for the past 40 years [1]–[3]. Moreover, this work has still faced many challenges because of limitation of data, various viewpoints, different scales, illumination conditions, complex background, and various modalities. To improve efficiency of action recognition results, some researchers try to increase the number of data using generative adversarial network (GAN) models such as [4]–[6]. Synthetic human action images are created by generator of GAN models which are similar to the training videos. Doan and Nguyen [7] generated hand images in multiple views with blender-based and hand glove-based. Although this dataset is diverse in viewpoints and variety of samples, it only provided static hand gestures.

Another approach could ameliorate action recognition results that utilizes multi-view cameras. Tran *et al.* [1] found discriminant of pairwise covariance of multiple views data to deal with the robustness of HAR result. This research finds transformation of multi-view actions in a common space. Then, a new action could be projected into the learned common space. But this approach composes of discrete blocks and it is difficult to deploy an end-to-end solution. Nguyen and Nguyen [8] proposed a dynamic action recognition method with residual network (ResNet)18 backbone, (2+1)D architecture, cross view attention (CVA) module and augmentation strategy. This work improved action recognition that used image sequences in multiple view-points but this method also required many cameras in both training and testing phases. An end-to-end HAR solution at an arbitrary viewpoint is necessary to deploy a real HAR application because of a simpler testing environment setup. Zhang *et al.* [9] composed a view-invariant transfer dictionary and classifier for novel-view action recognition. Two-dimensional (2D) videos are projected into a view-invariant sparse representation. Dictionary learning projection is considered as a linear algorithm that is quite a limitation. Gedamu *et al.* [10] proposed method to recognize action in a certain view but this solution is implemented by skeleton image and recognized static action.

Stimulated from Tran *et al.* [5], Doan *et al.* [11] proposed an end-to-end framework for an arbitrary view dynamic action recognition that combined the ArVi-MoCoGAN model, 3D convolutional (C3D) block and attention module. Both generator and discriminator of ArVi-MoCoGAN are utilized on testing phase to create multi-view synthetic actions which could increase the computational complexity of the system. Furthermore, in this research, C3D is used as a 3D feature extractor of multi-view synthetic video. They are then used as inputs of the attention module to vote channel attention and create the final feature vector before passing the soft-max layer. This attention module is not observed for spatial features. In this work we propose a new framework for an arbitrary view HAR that deals not only the channel attention but also the spatial attention. In addition, this work also investigates and compares on various 3D CNN extractors (3 dimension convolutional neural network).

In general, our research composes of two contributions, such as: (i) we propose a new arbitrary view gesture recognition method; (ii) investigate the arbitrary HAR framework with various 3D CNN extractor backbones. The remainder of this research is organized as follows: firstly, section 2 explains our proposed framework. Next, the experimental results are analyzed and discussed in section 3. Finally, section 4 consists of the conclusion of research direction as well as its future works.

## 2. PROPOSE METHOD

Our proposed dynamic action recognition method in certain unknown viewpoints is illustrated in Figure 1 that consists of four cascade main blocks: (i) generate the synthetic videos from a certain real video with ArVi-MoCoGAN model in [11]; (ii) feature extraction of the synthetic videos using various 3D CNN models; (iii) finding attentions of channels/viewpoints and spatial with convolutional block attention module (CBAM) [12]; and (iv) classification. Our framework is explained in the next parts from section 2.1 to section 2.4. In addition, section 2.5 presents multiview datasets, protocol and setup parameters for the entire experiment.

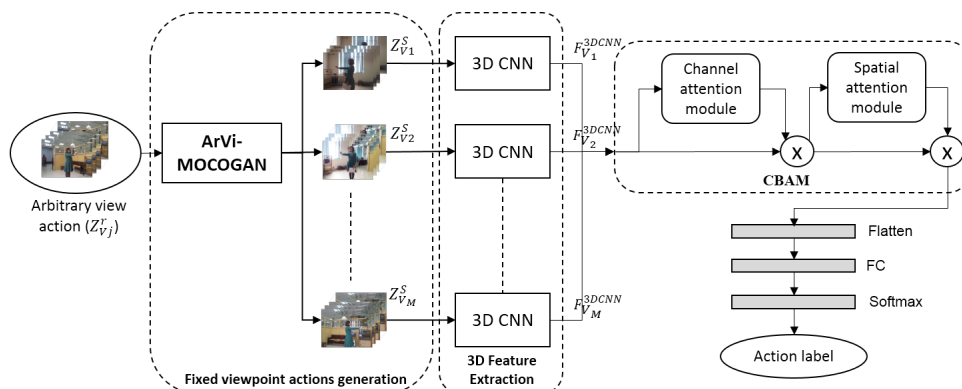


Figure 1. Framework of arbitrary view dynamic action recognition

## 2.1. Generation of the synthetic fixed videos

A common space is firstly created by ArVi-MoCoGAN model which is trained by the fixed real videos (M fixed cameras are setuped and captured for the training dataset). A novel real video is then projected in this common space to generate M synthetic videos in M fixed viewpoints. This architecture is similar to the ArVi-MoCoGAN model as presented in our previous research [11].

Given a real dynamic action in unknown view  $Z_{V_i}^r = [I_{V_i}^{(1)}, \dots, I_{V_i}^{(N)}]$ , N is number of frames in an arbitrary real video. Outputs of the trained ArVi-MoCoGAN model are the M synthetic videos  $\{Z_{V_j}^S = [I_{V_j}^{S,(0)}, \dots, I_{V_j}^{S,(N)}], j = (1, \dots, M), j\}$  on the M fixed viewpoints as illustrated in (1).

$$\mathcal{M}^{ArVi-MoCoGAN}(Z_{V_i}^r) = \{Z_{V_j}^S, j = (1, \dots, M)\} = \begin{cases} Z_{V_1}^S = [I_{V_1}^{S,(1)}, \dots, I_{V_1}^{S,(N)}] \\ Z_{V_2}^S = [I_{V_2}^{S,(1)}, \dots, I_{V_2}^{S,(N)}] \\ \dots \\ Z_{V_M}^S = [I_{V_M}^{S,(1)}, \dots, I_{V_M}^{S,(N)}] \end{cases} \quad (1)$$

where M equals to 5, 4, 7, and 3 that corresponds to the number of classes of MICAGes, IXMAS, MuHAVi and NUMA datasets respectively. Doan *et al.* [11] used both synthetic videos and view predicted probabilities of a novel real video on the fixed views for the HAR phase. In this work, only M synthetic videos of the ArVi-MoCoGAN model are utilized as inputs of the 3D CNN feature extractors in the next step.

## 2.2. 3DCNN feature extraction

Doan *et al.* [11] only used C3D network for feature extraction. In this research, four state-of-the-art (SOTA) 3D CNN models are utilized to compare efficient of our end-to-end HAR model that concludes C3D [13], ResNet50-3D [14], and RNN [15], [16], spatial-temporal attention [17]. These models are deployed as follows:

- C3D network is introduced in [13] with visual geometry group (VGG) as backbone that has become an efficient method for spatial and temporal 3D CNN method for action recognition. In this work, the 2048-D feature vector is extracted from FC7 layer. We utilize the network with batch normalization after every Conv layer, the pre-trained weights on Sports-1M and fine-tuned on Kinetics dataset [18].
- ResNet50-3D is built with ResNet50 backbone and 3D Conv layer. The spatial and temporal feature vector is taken after global average pooling layer whose dimension is 2048-D outputting. The Kinetics pre-trained weights are applied as in [18].
- ResNet50-TP (ResNet50 temporal attention) also uses ResNet50 backbone and TP. Output feature vectors of the ResNet50-TP model equals to 2048-D. The Kinetics pre-trained weights are applied as in [18].
- Recurrent neural network (RNN) architecture [19] and InceptionNetV3 model [20] are applied for dynamic action feature extractor. Dimension of feature vector is 512-D. Model is also fine-tuned by the Kinetics dataset [18] on all layers of this feature extractor.

In this paper, the 3D CNN model is used as spatial-temporal feature extractor. Inputs of the 3D CNN extractors are the fixed multi-view synthetic videos  $\{Z_{V_j}^S | j = (1, \dots, M)\}$  which are outputs of the previous ArVi-MoCoGAN model (in section .1). Outputs of the 3D CNN extractor are feature vectors  $\{F_{V_j}^{3DCNN} \in \mathbb{R}^{1 \times K} | j = (1, \dots, M)\}$  as illustrated in (2).

$$\mathcal{M}^{3DCNN}(Z_{V_j}^S, j = (1, \dots, M)) = \begin{cases} \mathcal{M}_{V_1}^{3DCNN}(Z_{V_1}^S) = F_{V_1}^{3DCNN}[1 \times K] = [F_{V_1}^{(0)}, \dots, F_{V_1}^{(K)}] \\ \mathcal{M}_{V_2}^{3DCNN}(Z_{V_2}^S) = F_{V_2}^{3DCNN}[1 \times K] = [F_{V_2}^{(0)}, \dots, F_{V_2}^{(K)}] \\ \dots \\ \mathcal{M}_{V_M}^{3DCNN}(Z_{V_M}^S) = F_{V_M}^{3DCNN}[1 \times K] = [F_{V_M}^{(0)}, \dots, F_{V_M}^{(K)}] \end{cases} \quad (2)$$

Where K equals 512 with the RNN feature extractor model and K is 2048 with C3D, ResNet50-3D, and ResNet50-TP feature extractor models.

### 2.3. Attention module

Feature vectors of synthetic videos  $\{Z_{V_j}^S | j = (1, \dots, M)\}$  on the multiple viewpoints ( $F^{3DCNN} = \{\mathcal{M}^{3DCNN}(Z_{V_j}^S) | j = (1, \dots, M)\} = \{F_{V_j}^{3DCNN} \in \mathfrak{R}^{1 \times K} | j = (1, \dots, M)\}$ ) are normalized and composed into  $F \in \mathfrak{R}^{M \times 1 \times K}$  as illustrated in (3).

$$F = [F_{V_1}^{3DCNN}[1 \times K], \dots, F_{V_M}^{3DCNN}[1 \times K]] = \begin{bmatrix} F_{V_1}^{(1)} & F_{V_2}^{(1)} & \dots & F_{V_M}^{(1)} \\ F_{V_1}^{(2)} & F_{V_2}^{(2)} & \dots & F_{V_M}^{(2)} \\ \dots & \dots & \dots & \dots \\ F_{V_1}^{(K)} & F_{V_2}^{(K)} & \dots & F_{V_M}^{(K)} \end{bmatrix} \quad (3)$$

Input of this module is  $F \in \mathfrak{R}^{M \times 1 \times K}$  where each feature vector element  $F_{V_j}^{3DCNN} \in \mathfrak{R}^{1 \times K}$  in  $F \in \mathfrak{R}^{M \times 1 \times K}$  is considered as a channel of the channel attention module. The channel attention module infers one 1-D channel attention map  $a_c \in \mathfrak{R}^{M \times 1 \times 1} = [a_c^{(1)}, a_c^{(2)}, \dots, a_c^{(M)}]$ . The output of the channel attention part  $F_c \in \mathfrak{R}^{1 \times K}$  is then calculated as illustrated in (4).

$$F_c = a_c \otimes F = \frac{\sum_{j=1}^M (a_c^{(j)} + 1) * F_{V_j}^{3DCNN}}{M} \quad (4)$$

Where  $\otimes$  denotes element-wise multiplication. Each the feature vector is paired with an attention values accordingly. It is then combined with itself which is copied along the spatial dimension.  $F_c \in \mathfrak{R}^{1 \times K}$  is passed over the spatial attention module. A spatial attention map  $a_s \in \mathfrak{R}^{1 \times 1 \times K}$  is also computed. It is then utilized to calculate the output of the CBAM model  $F_{CBAM} = F_s \in \mathfrak{R}^{1 \times K}$  that is presented as illustrated in (5).

$$F_{CBAM} = F_s = a_s \otimes F_c = a_s \otimes a_c \otimes F \quad (5)$$

### 2.4. Classification

Output feature vector ( $F_{CBAM}$ ) of the CBAM module as shown in (5) is flatten and fully connected together before being passed through a Softmax layer to classify. In this research, the Softmax cross-entropy loss function is applied to train and test entire networks. Given an arbitrary view dynamic action ( $Z_{V_i}, (i \neq j)$ ), its predicted result is  $\bar{p}_i$ . Its ground truth is  $p_i$ . Thus, the loss function is calculated as illustrated in (6).

$$L_{softmax} = \frac{1}{K} \sum_{i=1}^K p_i \log \bar{p}_i \quad (6)$$

### 2.5. Datasets, protocol and setup parameters

Dataset: in this research, four benchmark datasets are utilized, consisting of the MICAGes [21], IXMAS [3], MuHAVi [22], and NUMA [6] which contain 1,500, 1,584, 3,038, and 1,475 videos, respectively. They are the multiview dynamic action datasets which were mentioned in detail in [11]. Protocol: an arbitrary view evaluation protocol in [11] is utilized to test our framework on a single dataset. Where each view is separated and seen as an arbitrary viewpoint, remaining views are used as the fixed viewpoint. Testing is implemented by leave-one-view-out protocol entire viewpoints ( $V_j, j = (1, \dots, M)$ ) to achieve the final result.

Setup parameter: our model is deployed with two stages: firstly, generator and discriminator of an ArVi-MoCoGAN model are trained. Then, all layers of the generator of the ArVi-MoCoGAN model are used and frizzed in training of the arbitrary view dynamic action recognition framework as shown in Figure 1. Learning rate is  $5 * 10^{-5}$ ; optimizer is Adam; batch size equals 32 images; loss function is cross entropy; input image size is  $224 \times 224$  pixels. Quantitation results are compared in the next section 3.

## 3. EXPERIMENTAL RESULT

The evaluation schemes are written in Python on a Pytorch deep learning framework and run on a workstation with NVIDIA GPU 11G. The experiments are conducted to indicate the following problems: (i) comparison accuracy of our arbitrary view action recognition framework using various 3D-CNN backbones; (ii) parameters of various arbitrary view action models; and (iii) comparison of our best action recognition models with SOTA HAR methods.

### 3.1. Arbitrary view gesture recognition with various 3D CNN feature extractors

In this section, our novel view action recognition framework is evaluated by different 3D CNN backbones such as C3D, ResNet50-D, RNN, and ResNet50-TP. It is tested on various benchmark multi-view action datasets consisting of MICAGes, NUMA, IXMAS, and MuHAVi datasets. Results are presented in Figure 2.

It is evident that our arbitrary view action recognition framework with C3D backbone obtains the best accuracy on MICAGes, NUMA, IXMAS, and MuHAVi datasets at 96.6%, 92.79%, 93.01%, and 99.05% respectively. These percentage results are far higher than using the remaining 3D CNN feature extractors (ResNet50-D, RNN, and ResNet50-TP). While the RNN feature extractor has the lowest accuracy at 72.54% on MICAGes and 46.02% on NUMA; the ResNet50-TP backbone achieves the smallest accuracy at 65.5% on IXMAS and 80% on MuHAVi. Thus, the arbitrary view action recognition framework will be considered and compared by other factors in the next section 3.3.

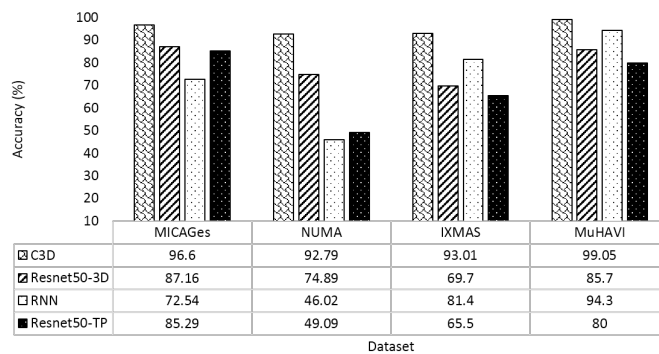


Figure 2. The accuracy of arbitrary view gesture recognition with various 3D CNN feature extractors

### 3.2. Summary of the arbitrary view dynamic action recognition models

This section summarizes in terms of params, FLOPs, time cost and model size of various arbitrary view dynamic HAR models with different 3D CNN backbones that is trained by MICAGes dataset as illustrated in Table 1. Where params represents the number of the trained model parameters. FLOPs shows the number of floating point operations required by the trained model. Time cost is the time total that the model processes from the beginning time to the ending time. Model size refers to the size of the container which contains the trained model on a certain dataset. Parameter calculation of the arbitrary view HAR system (Figure 1) is divided into two parts: Arvi-MoCoGAN model (second row of Table 1), 3D CNN and CBAM model (from third row to seventh row of Table 1). This table indicates some highlight issues that:

- The ArVi-MoCoGAN model has 10.8 (M) in params, 90.75 (G) in Flops, 0.218 (s) in time cost, and 40.39 (MB) in model size. These results indicate that params number and model size of ArVi-MoCoGAN model are smaller but time cost and FLOPs are higher than remaining parts of the end-to-end HAR framework.
- By comparing between 3D CNN extractors and CBAM model, it is apparent that using C3D + CBAM (Ford column and third row of Table 1) has the smallest time cost at only 0.105 (s) that is dramatically smaller than 1.77 (s) of ResNet50-3D + CBAM, (0.211 (s) of ResNet50-TP + CBAM and 0.171 (s) of RNN + CBAM. While params, FLOPs and model size are larger the remaining 3D CNN backbones at 73.37 (M), 38.70 (G), and 293.52 (MB) respectively. Despite using C3D extractor has high params, FLOPs and model size but it is the smallest time cost and the best HAR accuracy (section 3.1). Thus, this is worth attention and trade-off problems for a real application.

As a result, our end-to-end arbitrary view HAR system using C3D has a time cost total of 0.323 (s) that is equivalent to 3(fps) while it obtains the best accuracy at 96.6% on MICAGes dataset. These results can be accepted in order to deploy a real application.

### 3.3. Comparison with the SOTA arbitrary view gesture recognition

In this section, we compare our best accuracy results with some SOTA methods on four benchmark datasets as illustrated in Table 2. A glance at Table 2 it is evident that our method obtains the higher accuracy on three published datasets than recent HAR methods, such as: 93.01% on IXMAS is larger than 87.25% in

[11], 79.4% in [23] and 79.9% in [24]. On MICAGes dataset, our method accounts 96.6% that is better than [8], [11], [25] from 3.72% to 7.89% in accuracy. On the MuHAVI dataset, our approach also obtains the largest accuracy at 99.05%, it is higher than 0.78% in [11] and [23] at 5.45%. Our accuracy achieves 92.79% that is slightly smaller than [11] and [23] at 1.72% and 1.02% on NUMA dataset while it is dramatically better than the remaining methods in [8], [26]–[28] from 0.01% to 9.59%. This result once again indicates that our proposed solution is more efficient than recent methods in dynamic action recognition accuracy.

Table 1. Parameters of an arbitrary view dynamic action recognition model is trained by MICAGes dataset

	Params (M)	FLOPs (G)	Time cost (s)	Model size (MB)
ArVi-MoCoGAN	10.08	90.75	0.218	40.39
C3D + CBAM	73.37	38.70	<b>0.105</b>	293.52
ResNet50-3d + CBAM	55.43	10.15	0.177	222.04
ResNet50-TP + CBAM	23.55	17.39	0.211	94.51
RNN + CBAM	28.78	17.47	0.171	115.38

Table 2. Comparison of arbitrary view action recognition accuracy (%) using SOTA methods

	IXMAS	MICAGes	MuHAVI	NUMA
WLE [24]	79.9	-	-	-
SAM [26]	-	-	-	83.2
TSN [29]	-	-	-	90.3
DA-Net [27]	-	-	-	92.1
Multi-Br TSN-GRU [25]	-	88.71	-	93.81
R34(2+1)D With CVA [8]	-	91.71	-	92.78
$D_A + ELM + aug$ [23]	79.4	-	93.6	-
ViewCon + MOCO v2 [28]	-	-	-	91.7
ArVi-MoCoGAN + C3D [11]	87.25	92.88	98.27	<b>94.51</b>
<b>Our</b>	<b>93.01</b>	<b>96.60</b>	<b>99.05</b>	92.79

#### 4. CONCLUSION

In this research, a new arbitrary view HAR framework is proposed which combines a cascade blocks including an ArVi-MoCoGAN network, 3D CNN feature extractors and CBAM unit. Our method is deployed and evaluated by various 3D CNN models such as: C3D, ResNet50-3D, ResNet50-TP, and RNN. Our experimental result is implemented on different benchmark datasets. It shows that using C3D backbone obtains the best accuracy. In addition, our proposed framework archives higher efficiency than SOTA novel view action recognition on most benchmark datasets up to 9.59%.




#### REFERENCES

- [1] H.-N. Tran, H.-Q. Nguyen, H.-G. Doan, T.-H. Tran, T.-L. Le, and H. Vu, "Pairwise-covariance multi-view discriminant analysis for robust cross-view human action recognition," *IEEE Access*, vol. 9, pp. 76097–76111, 2021, doi: 10.1109/ACCESS.2021.3082142.
- [2] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, "Hand gesture recognition with 3D convolutional neural networks," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 1–7, doi: 10.1109/CVPRW.2015.7301342.
- [3] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding*, vol. 104, no. 2–3, pp. 249–257, 2006, doi: 10.1016/j.cviu.2006.07.013.
- [4] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "MoCoGAN: decomposing motion and content for video generation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 1526–1535, doi: 10.1109/CVPR.2018.00165.
- [5] T. H. Tran, V. D. Bach, and H. G. Doan, "vi-MoCoGAN: a variant of MoCoGAN for video generation of human hand gestures under different viewpoints," in *Proceedings of the Pattern Recognition: ACPR*, 2020, vol. 1180 CCIS, pp. 110–123, doi: 10.1007/978-981-15-3651-9\_11.
- [6] L. Wang, Z. Ding, Z. Tao, Y. Liu, and Y. Fu, "Generative multi-view human action recognition," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 6211–6220, doi: 10.1109/ICCV.2019.00631.
- [7] H.-G. Doan and N.-T. Nguyen, "New blender-based augmentation method with quantitative evaluation of CNNs for hand gesture recognition," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 30, no. 2, pp. 796–806, May 2023, doi: 10.11591/ijeecs.v30.i2.pp796-806.
- [8] H.-T. Nguyen and T.-O. Nguyen, "Attention-based network for effective action recognition from multi-view video," *Procedia Computer Science*, vol. 192, pp. 971–980, 2021, doi: 10.1016/j.procs.2021.08.100.
- [9] J. Zhang, H. P. H. Shum, J. Han, and L. Shao, "Action recognition from arbitrary views using transferable dictionary learning," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 4709–4723, Oct. 2018, doi: 10.1109/TIP.2018.2836323.




- [10] K. Gedamu, Y. Ji, Y. Yang, L. Gao, and H. T. Shen, "Arbitrary-view human action recognition via novel-view action generation," *Pattern Recognition*, vol. 118, p. 108043, Oct. 2021, doi: 10.1016/j.patcog.2021.108043.
- [11] H.-G. Doan, H.-Q. Luong, and T. T. T. Pham, "An end-to-end model of ArVi-MoCoGAN and C3D with attention unit for arbitrary-view dynamic gesture recognition," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 3, 2024, doi: 10.14569/IJACSA.2024.01503122.
- [12] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19, doi: 10.1007/978-3-030-01234-2\_1.
- [13] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 4489–4497, doi: 10.1109/ICCV.2015.510.
- [14] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 6546–6555, doi: 10.1109/CVPR.2018.00685.
- [15] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997, doi: 10.1109/78.650093.
- [16] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3D human action recognition," in *Computer Vision–ECCV 2016: 14th European Conference*, 2016, pp. 816–833, doi: 10.1007/978-3-319-46487-9\_50.
- [17] J. Wang and X. Wen, "A spatio-temporal attention convolution block for action recognition," *Journal of Physics: Conference Series*, vol. 1651, no. 1, p. 012193, Nov. 2020, doi: 10.1088/1742-6596/1651/1/012193.
- [18] W. Kay *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017, [Online]. Available: <http://arxiv.org/abs/1705.06950>.
- [19] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," *arXiv preprint arXiv:1402.1128*, 2014, [Online]. Available: <http://arxiv.org/abs/1402.1128>.
- [20] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.
- [21] H. G. Doan *et al.*, "Multi-view discriminant analysis for dynamic hand gesture recognition," in *Pattern Recognition. ACPR 2019. Communications in Computer and Information Science*, 2020, pp. 196–210, doi: 10.1007/978-981-15-3651-9\_18.
- [22] F. Murtaza, M. H. Yousaf, and S. A. Velastin, "Multi-view human action recognition using 2D motion templates based on MHIs and their HOG description," *IET Computer Vision*, vol. 10, no. 7, pp. 758–767, Oct. 2016, doi: 10.1049/iet-cvi.2015.0416.
- [23] N. Nida, M. H. Yousaf, A. Irtaza, and S. A. Velastin, "Video augmentation technique for human action recognition using genetic algorithm," *ETRI Journal*, vol. 44, no. 2, pp. 327–338, Apr. 2022, doi: 10.4218/etrij.2019-0510.
- [24] J. Liu, M. Shah, B. Kuipers, and S. Savarese, "Cross-view action recognition via view knowledge transfer," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2011, pp. 3209–3216, doi: 10.1109/CVPR.2011.5995729.
- [25] A.-V. Bui and T.-O. Nguyen, "Multi-view human action recognition based on TSN architecture integrated with GRU," *Procedia Computer Science*, vol. 176, pp. 948–955, 2020, doi: 10.1016/j.procs.2020.09.090.
- [26] S. Mambou, O. Krejcar, K. Kuca, and A. Selamat, "Novel cross-view human action model recognition based on the powerful view-invariant features technique," *Future Internet*, vol. 10, no. 9, pp. 1–17, 2018, doi: 10.3390/fi10090089.
- [27] D. Wang, W. Ouyang, W. Li, and D. Xu, "Dividing and aggregating network for multi-view action recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 457–473, doi: 10.1007/978-3-030-01240-3\_28.
- [28] K. Shah, A. Shah, C. P. Lau, C. M. de Melo, and R. Chellapp, "Multi-view action recognition using contrastive learning," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Jan. 2023, pp. 3370–3380, doi: 10.1109/WACV56688.2023.00338.
- [29] L. Wang *et al.*, "Temporal segment networks: towards good practices for deep action recognition," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 20–36.

## BIOGRAPHIES OF AUTHORS



**Anh-Dung Ho**    received B.E. degree in Applied Mathematics and Informatics in 2001, M.E. in Computer Science in 2007, all from Hanoi University of Science and Technology, Ha Noi, Vietnam. He can be contacted at email: [dungha@eaut.edu.vn](mailto:dungha@eaut.edu.vn).



**Huong-Giang Doan**    received B.E. degree in Instrumentation and Industrial Informatics in 2003, M.E. in Instrumentation and Automatic Control System in 2006 and Ph.D. in Control Engineering and Automation in 2017, all from Hanoi University of Science and Technology, Ha Noi, Vietnam. She can be contacted at email: [giangdth@epu.edu.vn](mailto:giangdth@epu.edu.vn).