

Attention correlated appearance and motion feature followed temporal learning for activity recognition

Manh-Hung Ha^{1,2}, The-Anh Pham³, Dao Thi Thanh⁴, Van Luan Tran⁵

¹Faculty of Applied Sciences, International School, Vietnam National University, Hanoi, Vietnam

²Faculty of Electrical and Electronic Engineering, Phenikaa University, Hanoi, Vietnam

³Prover Technology AB, Stockholm, Sweden

⁴Department of Software Engineering, FPT University, Hanoi, Vietnam

⁵School of Engineering, Eastern International University, Binh Duong, Vietnam

Article Info

Article history:

Received Jan 12, 2022

Revised Oct 2, 2022

Accepted Oct 25, 2022

Keywords:

Activity recognition

Attention mechanism

Deep neural network

Recurrent neural network

Spatiotemporal

ABSTRACT

Recent advances in deep neural networks have been successfully demonstrated with fairly good accuracy for multi-class activity identification. However, existing methods have limitations in achieving complex spatial-temporal dependencies. In this work, we design two stream fusion attention (2SFA) connected to a temporal bidirectional gated recurrent unit (GRU) one-layer model and classified by prediction voting classifier (PVC) to recognize the action in a video. Particularly in the proposed deep neural network (DNN), we present 2SFA for capturing appearance information from red green blue (RGB) and motion from optical flow, where both streams are correlated by proposed fusion attention (FA) as the input of a temporal network. On the other hand, the temporal network with a bi-directional temporal layer using a GRU single layer is preferred for temporal understanding because it yields practical merits against six topologies of temporal networks in the UCF101 dataset. Meanwhile, the new proposed classifier scheme called PVC employs multiple nearest class mean (NCM) and the SoftMax function to yield multiple features outputted from temporal networks, and then votes their properties for high-performance classifications. The experiments achieve the best average accuracy of 70.8% in HMDB51 and 91.9%, the second best in UCF101 in terms of 2DConvNet for action recognition.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Manh-Hung Ha

Faculty of Applied Sciences, International School, Vietnam National University

Hanoi 100000, Vietnam

Email: hungm@vnu.edu.vn

1. INTRODUCTION

Convolutional neural networks (CNNs) with fixed-size input and output vectors at sophisticated structures have been successfully demonstrated with high performance in the classification of subject activities in videos [1]–[3]. In order to improve performance in visual perception, several generations of CNNs have been created with the input vectors taking care of one image or multiple images. Particularly, multiple images are commonly adopted as an input vector which has the embedded temporal information as well as the spatial information [2], [4]. In addition to improving learning, many researchers used temporal networks to perform large-scale visual learning and activity classification from video clips, where temporal networks had recurrent connections to aid in video context understanding regarding time [2], [4]–[7].

Since all spatiotemporal patterns of subject interaction are not equally likely, some actions can be identified by their appearance alone [5]. For example, while playing musical instruments does not change

many motions in the video sequences, blowing candles is caused by the varying candlelight illumination, and the swing of a golf club is rapid. The recognition of the category can be handled by the semantics feature with a few red green blue (RGB) frames of video sequences. In contrast, the activity of walking, the transition from standing to walking, and running do not change the identity. The motion being performed can be at a fast-refreshing speed, and individual frames can be ambiguous. Therefore, motion cues provide a necessary approach by allowing the compensated optical flows to pick up potential [2], [4].

Another important reason is that current CNNs architectures are not able to take full advantage of temporal information and their performance is consequently often dominated by appearance recognition. The structure associated with temporal information plays a critical role in achieving good performance in activity recognition. Accordingly, we investigate a general temporal network structure that has a feature generation layer, a temporal layer, and a fully connected layer. Six topologies of many-to-one, many-to-many plus global maximum pooling, many-to-many plus global average pooling, many-to-many plus many-to-one, bidirectional many-to-one, and many-to-many plus bidirectional many-to-one associated with the temporal layer are further explored, along with two cells of long short-term memory (LSTM) and gated recurrent unit (GRU). Instead of the usual feed-forward neural network with dropout and softmax making the final prediction, in the inference classifier, the predicting voting classifier (PVC) scheme based on the multiple nearest class mean (NCM) classifiers [8], SoftMax, and majority voting [9] are developed to determine the action class. In this study, two datasets from HMDB51 and UCF101 [2] are adopted to evaluate the proposed deep neural network (DNN). To ensure a fair comparison, the UCF-101 dataset is used to validate the temporal network structures of 12 different types, which are then simulated and compared to the best one. Simulation results reveal that the temporal network structure using the bi-directional GRU layer yields the best performance, with an average accuracy of around 91.8% (split 1). It is because neighboring image frames in a video clip have forward and backward relationships. Additionally, GRU may be superior to LSTM under some conditions. As compared to the conventional DNNs, the proposed temporal network using the bi-directional GRU layer is fairly good for realizing activity recognition.

The deep learning architecture approaches can learn some representational features automatically, and their impressive results lead to the extensive use of them in various pattern recognition domains. Xception [10] is an extension of the inception architecture that replaces the standard Inception modules with depth-separable convolutions. It is applied in 2D space and has been proven to be powerful in terms of extracting spatial information [9]. Some action recognition studies [11]–[14] used two-dimensional convolutional neural networks (2DCNN) to extract spatial information, which is known as an auxiliary clue. In order to improve the accuracy, extending the CNN from image to video is the exploitation of temporal information. among various temporal network architectures, LSTM is the most popular one as it is able to maintain observations in memory for extended periods of time [15]. Further research explicitly demonstrated the robustness of LSTM even as experimental conditions deteriorated and indicated its potential for robust real-world recognition [16], [17].

To reliably and precisely generate subject descriptors, the recognition process may focus on the meaningful parts to increase the accuracy. For example, attention features were generated automatically from the DNN's intermediate layer(s) and then used to focus on the most meaningful part of an image for identification [2], [4]. In [5], the recurrent mechanism that assigned the weighted attention to the feature map from the convolutional layer was proposed for action recognition based on RGB images. Instead of using the RGB stream, the spatiotemporal attention mechanism adopts the joint points from the 3D skeleton for action recognition. They developed an end-to-end network with three temporal networks that individually performed the classifications, by selectively focusing on the discriminative joints of the skeleton (spatial attention), and assigning weights to the key sequential images (temporal attention) [2], [4].

In the processing pipeline approach, the RGB stream and flow stream are applied to process the multiple streams of the video data. Two frames are sampled from each input video. The network processes the sequence frame and the predictions are merged simply by late fusion [18]. The fusion modes of early, late, average, concatenation, sum, and 3DCovNet are approached at several levels feature [2]. We now have predictions from the two streams, based on the spatial and temporal streams separately. The last step is to combine the two streams to produce the final output through an attention mechanism for the temporal network.

To effectively determine the subject's actions, many methods, such as Bayesian, hidden Markov models, gaussian mixture models (GMMs) [19], support vector machines (SVMs) [20], and feed-forward neural networks, are commonly used. To handle diversely growing data sets, it may be a good choice to use a model-free method. The classic model-free methods include the K-nearest neighbor (KNN) [21], and NCM classifiers [8]. The class-incremental learning mechanism was developed to train multiple NCM classifiers accompanied by feature representations simultaneously. The deep NCM (DNM) classifier directly learns highly non-linear visual representations to yield performance as good as the softmax optimized networks. The few-shot learning approach of the networks was used to learn a deep representation based on the NCM

classifiers [8]. Our contribution to this paper is as follows: i) we proposed an architecture for video activity recognition that is able to be composed using both rich spatial and temporal feature abstraction by attention mechanisms. This presentation enhances performance, enabling easier learning, and interpretability of the model; ii) we conducted extensive experiments on temporal networks and compared the six topologies of LSTM and GRU to obtain well-defined architectures in the UFC101 and HMDB51 datasets; and iii) we proposed the PVC method for incremental predicting, while achieving significantly better accuracy than existing incremental counterparts.

The remainder of this paper is organized as follows. Section 2, the proposed framework for a better solution is described in detail. We then conducted an implementation on the UCF101 and HMDB51 datasets. Section 3 shows the effectiveness of the architecture, and an analysis of the obtained result. Finally, the paper is concluded in section 4.

2. PROPOSED DNN FOR ACTION RECOGNITION

As shown in Figure 1, the genetically proposed DNN consists of two stream fusion attention (2SFA), a temporal layer, and a classifier is devised for action recognition. In the first step, a video clip is decomposed into multiple video segments. Each of these has an interval of a few seconds that is sufficient to contain an action. The neighboring video segments are overlapped so that the RGB and optical flow sequences from each video segment are used as the inputs. Owing to the fixed dimension of the input neural layer in the proposed DNN, the number of frames and frame size in a video segment may need to be converted. The preprocessing techniques of upsampling, downsampling, and size scaling are employed to transform frames in a video segment into the required number of specific size frames, which are inputted to two 2DCNNs to produce appearance and motion feature maps. In each video segment, the fusion attention (FA) generation layers yield the local descriptors. The temporal networks continuously process the outputs of the FA generation layers to generate latent spatial-temporal features. Finally, the inference classifier of the PVC scheme is utilized to attain the final class determination.

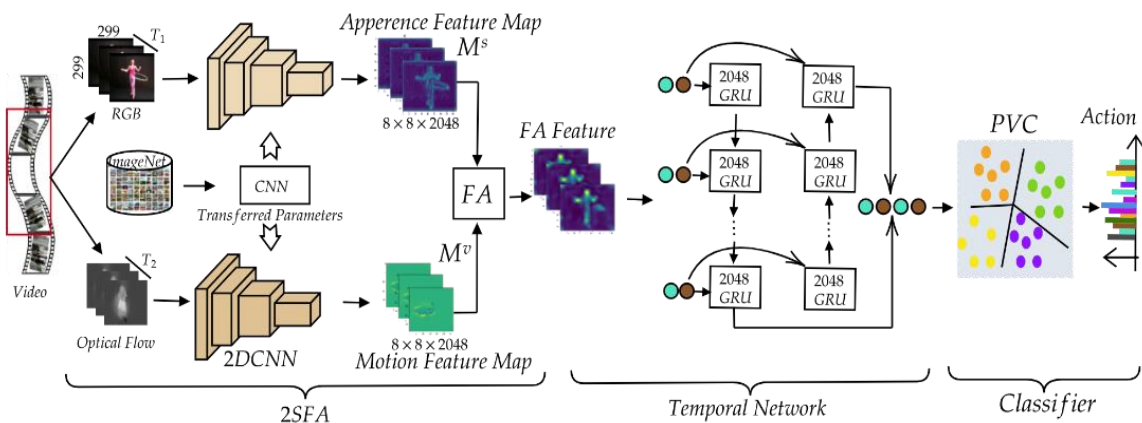


Figure 1. Our proposed DNN including 2SFA followed temporal network for action recognition

2.1. 2SFA network

Figure 1 illustrates our concept of a generic network which can be described in parallel, namely appearance feature and motion feature fusion based attention. The output of the fusion of two streams by the attention mechanism will be fed into the bidirectional GRU one layer to learn temporal information. Our architecture is empirically more effective for video recognition.

2.1.1. Appearance RGB and flow stream adopt backbone 2DCNN structure

For the RGB stream, work on a video clip that captures the spatial appearance feature. T individual video frames were used as network inputs, followed by several convolutional layers, pooling layers on pre-trained ImageNet on the Inception V3 model [2], and finally fine tuning the UCF101 and HMDB51 datasets. Finally, after fully connected (FC) layers, the network outputs are taken as the predicted probabilities of the video classes by the FA layer to yield probability.

In order to capture the motion information, the sampled $k*T$ frame is used for the flow stream. In our experiment, the value of $k=6$ for HMDB51 and $k=5$ for the UCF101 dataset was set. The temporal 2DCNN takes stacked optical flows as input of a pair of consecutive frames, the horizontal and vertical components of the calculated displacement vector fields. To further consider temporal information, one can stack the optical flow images of each frame at time t and its subsequent frames into a stacked 2L-channel optical flow image [2]. The network architecture and training process of the temporal CNN are basically the same as their spatial counterparts, except that the input images have a different number of channels. There are multiple stacked 2L-channel optical flow images in the video. The way of fusing predictions on these individual images is also the same as that of the spatial channel. We now have predictions from the two CNNs, based on the spatial and temporal streams separately. The last step is to combine the two streams to produce the final output.

2.1.2. FA layer scheme

In order to be sensitive to the meaningful details of an action, we propose the FA generation layers between the 2DCNNs and temporal networks by highlighting the distinguishable components among actions. The FA generation layer, as depicted in Figure 2, combines the appearance feature map with the coordinate values of the motion feature map and computes the combined data to yield the attention weights. After that, it convolutes the attention weights with the feature element to produce the output result of the FA. It is because the tracks of some actions depend on a part of a subject's movement. Hence, an attention feature map with reasonably emphasized regions can be beneficial to action distinction.

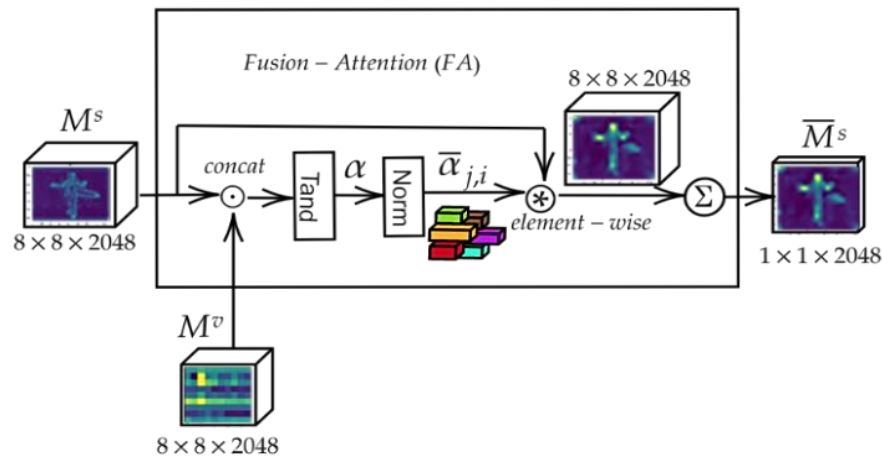


Figure 2. The proposed fusion attention for correlation feature

As shown in Figure 2, the operations of the FA generation layer are formulated as follows. M^s denotes the appearance feature map with the dimension of $k \times k \times D$. M^v represents the coordinate values of motion feature map from the RGB frames, with the same dimension. The first operation is to cascade M^s and M^v to become an input vector which addresses the fully connection layer with the activation function of tanh to generate the $k \times k \times D$ output, α , by

$$\alpha = F_c(M^s \odot M^v) \quad (1)$$

where the function $F_c(\cdot)$ denotes to the fully convolution, and \odot refers to the concatenated operation. This full connection layer performs the operations of the input vector multiplied by the corresponding weights, and added with the biases to become the accumulated data, which go through the tanh function to yield output results. The abovementioned operations are to build up the correlation between the appearance feature map and the motion feature map. Second, the attention parameters α are normalized according to a single feature frame at the dimension of $k \times k$ in (2),

$$\bar{\alpha}_{j,i} = \frac{\exp(\alpha_{j,i})}{\sum_{r=1}^{k \times k} \exp(\alpha_{j,r})} \quad (2)$$

where $\bar{\alpha}_{j,i}$ represents the normalized attention weight at the i^{th} component of the r^{th} feature frame, and $\alpha = \{\alpha_{j,i} | j = 1, \dots, D; i = 1, \dots, k^2\}$. This operation somehow exponentially increases the impact of attentions with positive values and normalize them to show out the relatively important ones at each feature frame. Finally, the normalized attention weights are multiplied with the corresponding components of each single feature map by (3),

$$\bar{M}_j^s = \sum_{i=1}^{k \times k} \bar{\alpha}_{j,i} M_{j,i}^s, \tag{3}$$

where $\bar{M}^s = \{\bar{M}_j^s | j = 1 \dots D\}$ is the output, the attention feature of the FA generation, such an operation embeds normalized attentions to the static feature map to highlight the critical part at each single $k \times k$ feature map as well as reduces the dimension.

2.2. Different temporal network for activity recognition

In this study, we introduce the temporal network for enhancement classification of the temporal sequence feature. The temporal network in Figure 1 is depicted in Figure 3 with six temporal topologies, which are many-to-one in Figure 3(a), many-to-many plus global maximum pooling in Figure 3(b), many-to-many plus global average pooling in Figure 3(c), many-to-many plus many-to-one in Figure 3(d), bidirectional many-to-one in Figure 3(e), and many-to-many plus bidirectional many-to-one to interpret the spatial and temporal relationship in Figure 3(f). The many-to-one network has T cells to yield one output vector. Each topology of many-to-many plus global maximum and average pooling has one output vector generated from global maximum and average pooling, respectively, where each of them includes T cells. The network of many-to-many plus many-to-one is to cascade the many-to-many and many-to-one layers where 2T cells are used to produce one output vector. The bi-directional many-to-one network consists of 2T cells with forward and backward connections to generate two output vectors. The network of many-to-many plus bi-directional many-to-one needs 3T cells to bring out two output vectors. The bidirectional networks not only provide the memories for learning dependencies but also allow the networks to make predictions from the future to the past as well as from the past to the future, in order to increase the classification power of the network layer.

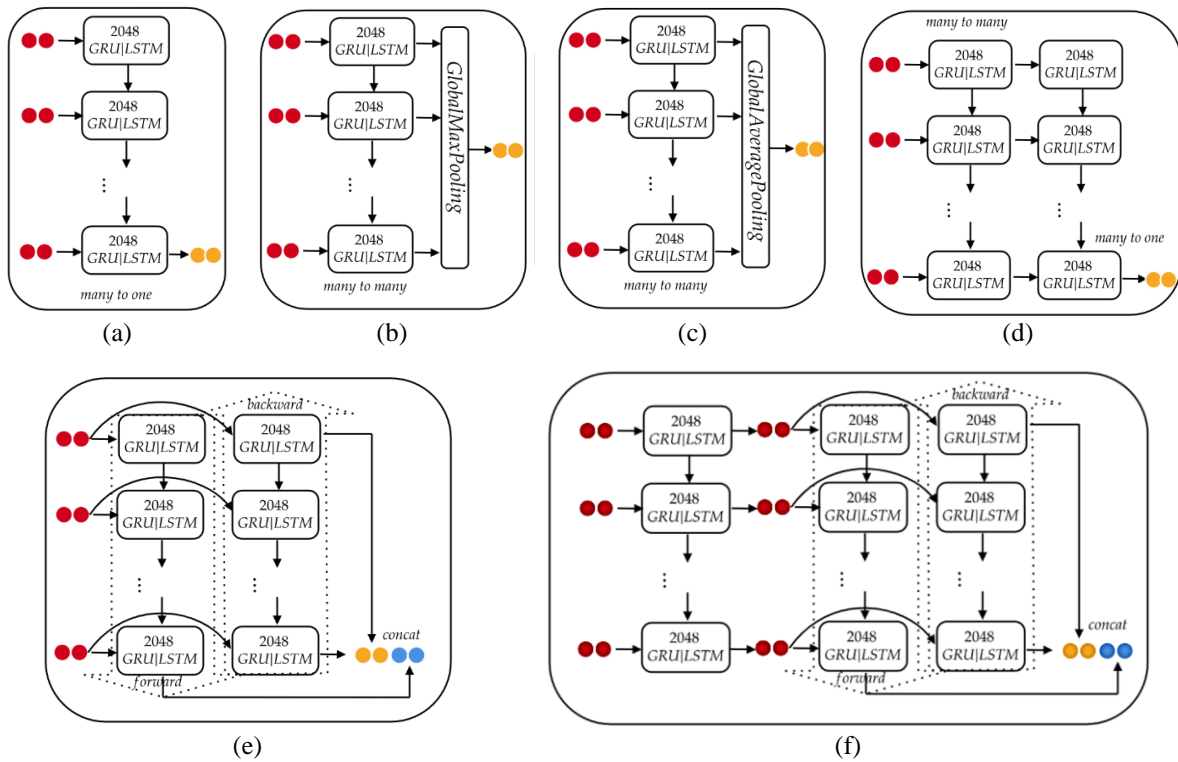


Figure 3. Six temporal topologies: (a) many-to-one, (b) many-to-many plus global maximum pooling, (c) many-to-many plus global average pooling, (d) many-to-many plus many-to-one, (e) bidirectional many-to-one, and (f) many-to-many plus bidirectional many-to-one

The cells inside these topologies have two choices of LSTM and GRU that are able to keep the memory and state and to remember the characteristics from the previous cell in the long term. LSTM discloses the memory state unit with separate input and forget gates, whereas GRU exposes whole state information to the other cells through its reset gate. The structure of LSTM includes nonlinear function gates and a memory cell. The information of the next memory cell is adjusted based on the previous cell's memory, and the input gate is activated by multiplying the activation from the forget gate and the signal from the input gate. LSTM also utilizes the output gate to control the information received by a hidden state variable. Similarly, GRU has gating units that modulate the information flow inside the cell without a separate memory unit. The main differences between LSTM and GRU can be described as follows: In GRU, a single gating unit simultaneously controls the forgetting factor and the decision to update the state unit. The reset and update gates can individually ignore a part of the state vector. The update gates act like conditional leaky integrators that can linearly gate any dimension, thus choosing to copy or completely ignore it by replacing it with the new target value. The reset gate controls which parts of the state are used to compute the next target state, introducing an additional non-linear effect in the relationship between the past and future states [21].

2.3. Incremental predicting class

In Figure 1, the outputs from temporal networks pass through the individual fully-connected neural layers to yield feature vectors for further classification. Based on these two feature vectors, the classifiers of SVM, GMM, DNCM, KNN, global average pooling (GAP)+SoftMax, FC+SoftMax, and PVC are investigated and compared. In this work, the PVC scheme, including the NCM classifier, SoftMax, and majority voting, is developed to achieve good performance.

The NCM classifier can be regarded as incremental class learning and classification [8]. The feature vector can be represented as an equally long sub feature vector of the corresponding class. Figure 4 shows the block diagram of the proposed PVC scheme, in which the output is the feature vector from the GRU by FC layer. Here, is a C-action capsule associated with the corresponding classes in a certain dataset which performs the NCM classifier to obtain its class group with the minimum Euclidean distance from the mean of that group.

In the beginning, the training data in each epoch is used to generate the class mean of the NCM classifier. Afterward, the class means are updated from the classifieds in the class groups at each epoch (each mini batch) [8]. The softmax fulfills the Euclidean distances from all feature vectors to compute the probability of each class to which the closest distance between feature vector and class mean may belong. Finally, the majority voting module from the scikit learn toolbox predicts the class label based on the argmax of the sum of the predicted probabilities to make the final class determination in a video clip. PVC will assign the input data to the class label with the largest probability by the argmax function. Unlike the loss function that is calculated by margin loss, during the training process, the loss function of PVC, which contains the recomputed mean of class at each epoch, is estimated by minimizing the cross-entropy loss as in a regular classification neural network in the similar formula of [8].

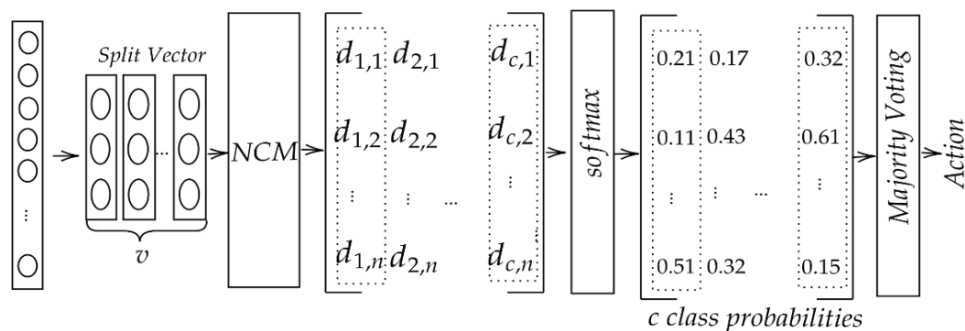


Figure 4. Proposed inference PVC

3. EXPERIMENT RESULTS AND DISCUSSION

In order to make sense of the sequence in our system, a single frame is used instead of a group of frames as a segment of the video. We assume that analyzing three seconds of video at a time is enough to make a good prediction of the activities. For this, the extracted frame of three seconds is downsampled to T frame (in this study, the typical value is T=15) into one single pattern, which will be the input of the neural network. The length of the video is a separate pattern (or clips).

3.1. Dataset and parameter setups for experiments

The UCF101 dataset with 13,320 video clips associated with 101 action categories was used. The video clips in this dataset are split into three groups for training and validation (9,537 entries), testing (3,783 entries). The HMDB51 dataset has 6,766 video clips concerning 51 action categories and diverse background contexts and variations in motion camera. There are three groups split into 3,570 training and validation and 1,530 testing data videos.

In each of the parameter setups and simulation situations, there are 100 epochs in each of which the well-trained model is obtained. When the accuracy cannot be incrementally improved at 10 sequential epochs, the training process is terminated by an early stopping scheme. Additionally, the validation loss and accuracy associated with the well-trained model at each epoch are stored. Among these values of validation loss and accuracy from all well-trained models, the model with the lowest loss and highest accuracy is chosen for the testing. At the training, the optimization is fulfilled by Adam. With a momentum of 0.9, the learning weights use mini-batch stochastic gradient descent. The number of frames extracted from a video clip is T , where T frames are input frames. Our computation platform employs an 8-core Intel Core i7 computer accompanied by a graphic processing unit (GPU) of a Nvidia Titan X 1080/12 GB with 32 GB RAM where the CPU system is the 64-bit Ubuntu 18.04, and the GPU is supported by the Anaconda distribution for Python.

3.2. Simulations and comparisons of temporal network at six type using LSTM and GRU

In this paper, we developed a model that combines CNN and the temporal network. In our model, a CNN is used to perform the task of extracting the features from each video. Besides the temporal network, we evaluated the action recognition framework by evaluating six temporal topologies using LSTM and GRU simulations. Finally, the classification layer using FC+softmax is used to predict the category. Figure 5 displays the validation loss and accuracy of the temporal networks at 12 types where the symbols of LSTM1/GRU1, LSTM-GMP/GRU-GMP, LSTM-GAP/GRU-GAP, LSTM2/GRU2, BDLSTM1/BDGRU1, BDLSTM2/BDGRU2, denote the temporal networks using LSTM/GRU at topologies of many-to-one, many-to-many plus global maximum pooling, many-to-many plus global average pooling, many-to-many plus many-to-one, bidirectional many-to-one, and many-to-many plus bi-directional many-to-one, respectively. In Figure 5(a), the performance of the temporal networks at six temporal topologies using LSTM is compared to conclude that BDLSTM1 yields the best. On the other hand, the performance of the temporal networks at the types of BDGRU1 exhibits the best, as depicted in Figure 5(b). The best one is BDGRU1 of the temporal networks at the bidirectional many-to-one topology using GRU cells, which performs well and is rapid. In particular, the temporal networks at the topologies using GRU converge faster than those at the corresponding topologies using LSTM. The accuracy of the temporal networks at the topologies using GRU is likely better than that of the corresponding topologies using LSTM at most categories of the UCF-101 dataset. GRU, on average, has lower complexity, easier modification, faster training, and higher performance than LSTM with fewer training data, whereas LSTM remembers longer sequences and outperforms at tasks requiring long-distance relationships. Table 1 compares the performances of the temporal networks at 12 types where F1 measure and testing accuracy are included, as well as training accuracy and validated accuracy/loss. The results reveal that 2SFA+BDGRU1 is the best, with a validation loss of 0.78 and a testing accuracy of 91.8%, and an F1 measure of 0.91 average micro and macro factor.

3.3. Model 2SFA + BGRU1 + PVC

Figures 6(a) and 6(b) also shows that the proposed model was learned on the UCF101 and HMDB51 datasets. Training validation and loss validation of 2SFA+BGRU1+PVC on UCF101 obtained an approximate 91% and 0.78, respectively. Performance on HMDB51 reached about 71% training validation, a 0.8 validate loss.

For quantitative evaluation, the PR curve is the better method performs as shown in Figure 6(c) and Figure 6(d) which depicted how the prediction-recall (PR) curve compare difference classification performance of GRU model. Figure 6(c), the 2SFA+BDGRU1 model achieved the best results on PR curves evaluation metrics. It means that on the UCF101 dataset, the proposed DNNs are better than the other models in terms of AUC-PR metrics. Figure 6(d), the AUC of the PR curve on the 2SFA+BDGRU1 model gets a promising result of 0.972 and 0.965 on UCF101 and HMDB51, respectively. We found that the proposed model can reach 1.0 for PR on UCF101, and 0.9 for HMDB51 respectively.

We compare two approach policies, 2SFA and 2SFA+BDGRU1, on the three different split ratios of the UCF101 and HMDB51 datasets. The results are shown in Table 2, which reveals the average accuracies of split 3-fold in the UCF101 and HMDB51 datasets. The experiments can reach 91.2% and 70.3% of 3-fold in the UCF101 and HMDB51 datasets, respectively, where the proposed system adopts 2SFA+BDGRU1. The results clearly show that the proposed system, which employs 2SFA and temporal bidirectional GRU in one layer, achieves the best performance. Most of our daily activities can be completed in this manner. In that

time interval, fusion RGB and flow stream architecture that is more likely closest to [2], [14] structure have 88.2 and 68.0 in both given datasets. By addition temporal bidirectional GRU one layer, our outcome comfortably outperforms them by a margin of 0.2% and 0.23 on UCF101 and HMDB51, respectively.

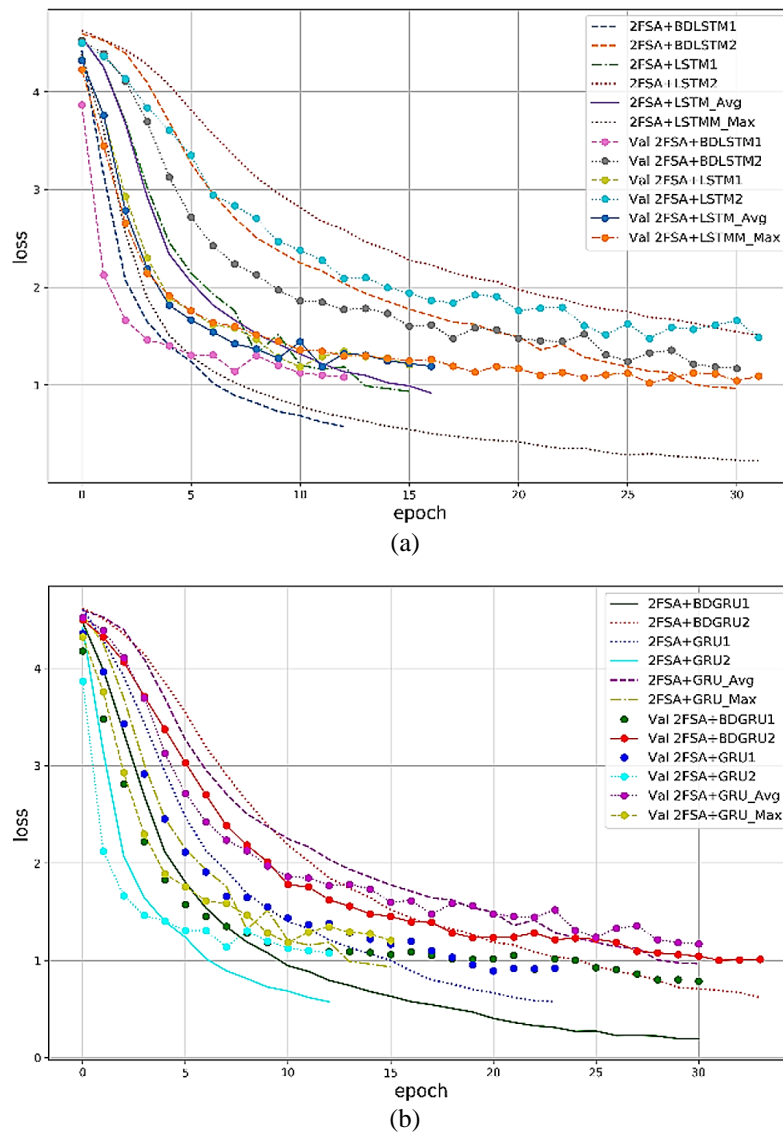


Figure 5. Validation loss of the 12 types of variant temporal networks, (a) evaluating six types of LSTM and (b) evaluating six types of GRU

Table 1. Performances of the proposed 2SFA + the temporal network (each 12 types) + FC+softmax on UCF101 dataset split 1

2SFA + Temporal Network Of:	Training performance				Average F1-Measure		Testing Acc.
	Train	ValTrain	Loss	ValLoss	macro	micro	
GRU1	98.7%	82.8%	0.44	0.92	0.83	0.82	83.1%
GRU-GMP	97.1%	77.1%	0.18	1.01	0.77	0.76	77.0%
GRU-GAP	96.7%	90.7%	0.49	0.92	0.88	0.89	89.4%
GRU2	88.8%	84.3%	0.84	1.07	0.81	0.82	82.8%
BDGRU1	99.6%	91.4%	0.12	0.78	0.92	0.90	91.8%
BDGRU2	90.7%	80.4%	0.59	1.00	0.82	0.79	81.2%
LSTM1	86.8%	83.6%	0.77	1.14	0.81	0.82	82.9%
LSTM-GMP	98.4%	79.9%	0.18	0.92	0.80	0.79	80.7%
LSTM-GAP	85.7%	73.5%	0.76	1.07	0.75	0.73	74.0%
LSTM2	66.1%	64.9%	1.16	1.48	0.64	0.66	65.6%
BDLSTM1	98.0%	84.2%	0.51	0.96	0.85	0.85	85.8%
BDLSTM2	87.0%	83.7%	0.93	1.14	0.79	0.82	81.1%

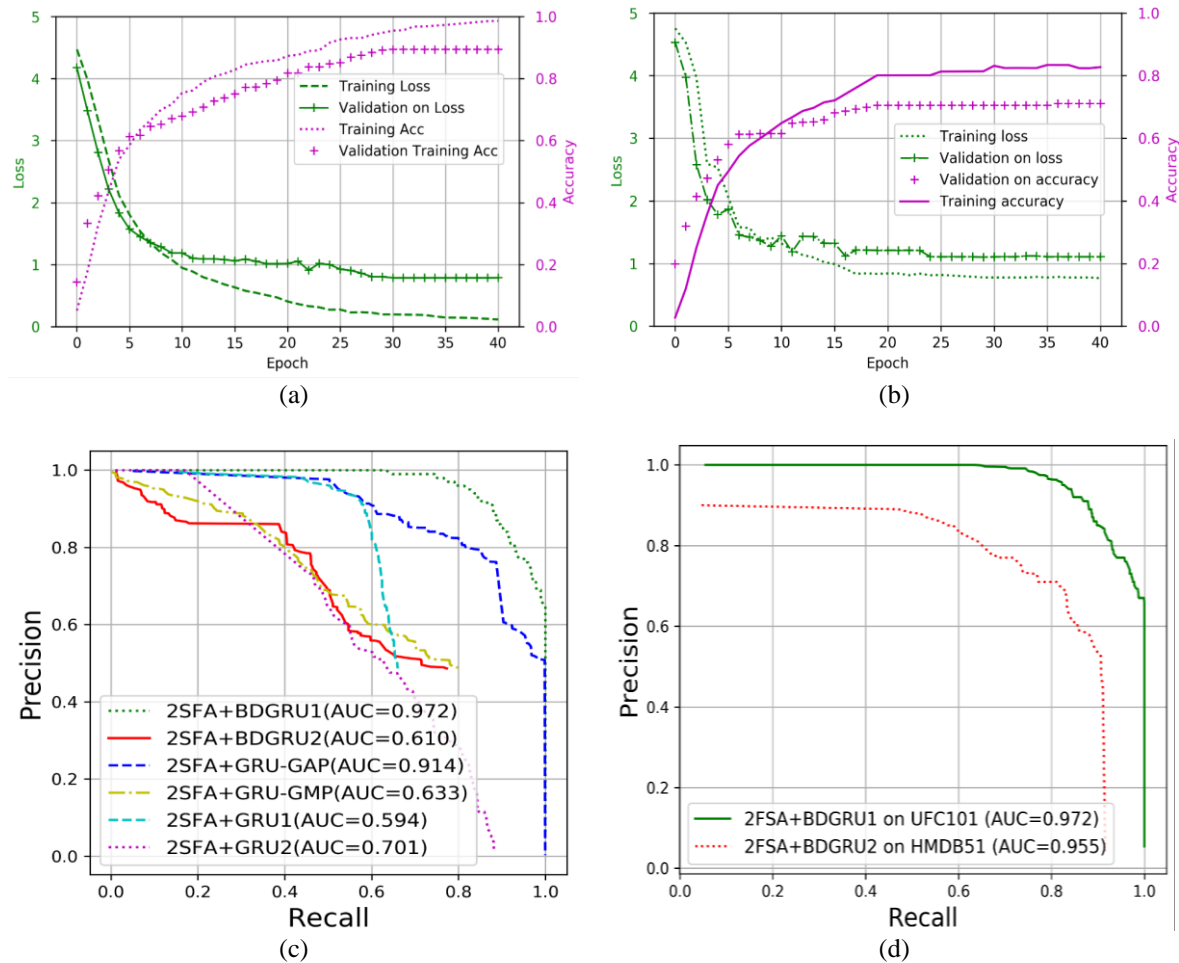


Figure 6. Model 2SFA + BGRU1 + PVC, (a) Training/Loss validation of 2SFA+BGRU1+PVC on UCF101, (b) HMDB51 dataset, (c) Precision – recall curves of six model on UCF101 dataset, and (d) Precision-recall curves of 2SFA + BGRU1 at the UCF101 and HMDB51

Table 2. Training/testing ratio

Data ratio	Classifications (%)	
	2SFA	2SFA +BDGRU1
UCF101 split 1	89.5	91.8
UCF101 split 2	88.1	91.4
UCF101 split 3	86.9	90.3
UCF101 3-fold	88.2	91.2
HMDB51 split 1	68.0	70.8
HMDB51 split 2	67.7	69.8
HMDB51 split 3	68.3	70.4
HMDB51 3-fold	68.0	70.3

Evaluation of Inference Classifiers: In addition to the feature generation, correlation, and highlights from the spatial and temporal domains, the final interpretation needs an adequate classifier in the proposed 2SCNN+FA+BDGRU1 to ameliorate the performance. Herein, the popularly used inference classifiers of FC+softmax, SVM, GMM, KNN, and our PVC are explored. The proposed PVC would work as a trainable feature distinguisher. Results listed in Table 3 at the UCF101 dataset indicate that the proposed PVC is superior to the other classifiers of FC+softmax, SVM, and GMM by 0.08%, 0.34%, and 1.07%, respectively, as well as increasing from 0.02% to 0.43% at the HMDB51 dataset. Hence, the proposed PVC is a good choice for the final classifier in our DNN.

We compare different fusion strategies in Table 4, where we report the average accuracy of the UCF101. We see that max and average perform considerably lower than sum and concatenation. For all the fusion methods shown in Table 4, fusion by our proposed FA results in higher performance compared to

other methods. The performance is slightly better because the layer spatial correspondences between appearance and motion are fused, which would have paid more attention by correlation attention weight in the area.

Table 3. Accuracy on test dataset. Comparison PVC to other method on multi-class classification (split 1)

	FC+ Softmax	SVM	GMM	PVC
UCF101	91.84	91.58	90.85	91.92
HMDB51	70.81	70.77	70.40	70.83

Table 4. 2SFA+BGRU1+PVC in difference fusion method

Fusion method	UCF101	HMDB51
Max	90.2	68.5
Average	90.5	68.9
Concatenation	90.9	69.0
Sum	91.3	69.2
FA	91.9	70.8

Table 5 shows the average accuracy of the comparison of our proposed to previous work DNNs in which all of the modalities are only using 2DConvNet at the same dataset. Because of the use of different network architectures and improvement schemes, the proposed DNNs are nearly as accurate as hybrid DNNs on UCF101 and achieve higher accuracy than those done on HMDB51 [9], [22], [23]. Most of the methods are not directly comparable to our results. We obtained a new state-of-the-art result of 70.8% and achieved a substantial high accuracy of 91.9% as second best by only using 2DConvNet.

Table 5. Performance comparison of the proposed model using only 2DConvNet classification on UCF-101 datasets

Modality	Feature Set	UCF101	HMDB51
ConvNet [11]	Slow fusion spatio-temporal	65.4	-
LRN [15]	Learning sequential dynamics	82.9	-
DT [12]	Multi-view super vector	83.5	55.9
LSTM-comp [16]	RGB+Flow model	84.3	-
iDT [24]	dense trajectories by camera motion	-	57.2
boVW [13]	Bag of visual words and fusion	87.9	61.9
FstCN [25]	SCI fusion	88.1	59.1
MoFAP [26]	Single Shot multi-Span in FC3D	88.3	61.7
Motion Infor [27]	TrajShape + TrajMF	78.5	57.0
	TrajShape + TrajMF + Wang and Schmid [18]	87.2	57.3
Gaussian Pyramid [22]	Multi-skip feat. stacking	89.1	65.4
Hybrid fusion+ DeepNet [23]	Supervised mid-to-end learning + non-linear classification	90.6	67.8
STAN [28]	FRA+OPF + CLIP	93.6	-
Confidence Distillation [14]	Distillation loss for student and teacher learning	91.2	-
2STG	2SFA+ BGRU1 + PVC	91.9	70.8




4. CONCLUSION

We have proposed a DNN architecture with a fusion attention layer, additional temporal networks, and a PVC layer. We investigate a general temporal network structure that has a feature generation layer, a temporal layer, and a fully-connected layer. The six topologies of the temporal layer are explored with the use of LSTM and GRU. The UCF-101 dataset was adopted for simulations of the temporal networks with the 12 resulting types. The experiment results reveal that the temporal networks at the bidirectional temporal layer using GRU show the best performance. The reason is that the bidirectional topology can effectively interpret the forward and backward temporal relationships in a video clip. Additionally, GRU exhibits relatively better performance than LSTM. Hence, the temporal networks at the bidirectional topology using GRU is recommended for the proposed fusion two stream flow temporal neural network. We also show that the proposed DNN model can improve performance with a classifier by PVC layer in the UCF101 and HMDB51 datasets, leading to the suggestion of the importance of learning on highly abstract spatial-temporal features. The simulation results show that the proposed system demonstrates fairly good performance in recognizing activities of subjects in different situations.




REFERENCES

- [1] H. Kim, S. Lee, and H. Jung, "Human activity recognition by using convolutional neural network," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 6, pp. 5270–5276, Dec. 2019, doi: 10.11591/ijece.v9i6.pp5270-5276.
- [2] M.-H. Ha and O. T.-C. Chen, "Deep neural networks using capsule networks and skeleton-based attentions for action recognition," *IEEE Access*, vol. 9, pp. 6164–6178, 2021, doi: 10.1109/ACCESS.2020.3048741.
- [3] M. A. Alsaedi, A. S. Mohialdeen, and B. M. Albaker, "Development of 3D convolutional neural network to recognize human activities using moderate computation machine," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 10, no. 6, pp. 3137–3146, Dec. 2021, doi: 10.11591/eei.v10i6.2802.
- [4] M.-H. Ha and O. T.-C. Chen, "Deep neural networks using residual fast-slow refined highway and global atomic spatial attention for action recognition and detection," *IEEE Access*, vol. 9, pp. 164887–164902, 2021, doi: 10.1109/ACCESS.2021.3134694.
- [5] M.-H. Ha and O. T.-C. Chen, "Action recognition improved by correlations and attention of subjects and scene," in *2021 International Conference on Visual Communications and Image Processing (VCIP)*, Dec. 2021, pp. 1–5, doi: 10.1109/VCIP53242.2021.9675340.
- [6] A. AL Smadi, A. Mehmood, A. Abugabah, E. Almekhlafi, and A. M. Al-smadi, "Deep convolutional neural network-based system for fish classification," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 2, pp. 2026–2039, Apr. 2022, doi: 10.11591/ijece.v12i2.pp2026-2039.
- [7] V. K. Kambala and H. Jonnadula, "A multi-task learning based hybrid prediction algorithm for privacy preserving human activity recognition framework," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 10, no. 6, pp. 3191–3201, Dec. 2021, doi: 10.11591/eei.v10i6.3204.
- [8] Y. Cheng, K.-Y. Wong, K. Hung, W. Li, Z. Li, and J. Zhang, "Deep nearest class mean model for incremental odor classification," *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 4, pp. 952–962, Apr. 2019, doi: 10.1109/TIM.2018.2863438.
- [9] O. T.-C. Chen, M.-H. Ha, and Y. L. Lee, "Computation-affordable recognition system for activity identification using a smart phone at home," in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, Oct. 2020, pp. 1–5, doi: 10.1109/ISCAS45731.2020.9180826.
- [10] F. Chollet, "Xception: deep learning with depthwise separable convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 1800–1807, doi: 10.1109/CVPR.2017.195.
- [11] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 1725–1732, doi: 10.1109/CVPR.2014.223.
- [12] Z. Cai, L. Wang, X. Peng, and Y. Qiao, "Multi-view super vector for action recognition," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 596–603, doi: 10.1109/CVPR.2014.83.
- [13] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," *Computer Vision and Image Understanding*, vol. 150, pp. 109–125, Sep. 2016, doi: 10.1016/j.cviu.2016.03.013.
- [14] S. M. Shalmani, F. Chiang, and R. Zheng, "Efficient action recognition using confidence distillation," *arXiv:2109.02137*, Sep. 2021.
- [15] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 677–691, Apr. 2017, doi: 10.1109/TPAMI.2016.2599174.
- [16] N. Srivastava, E. Mansimov, and R. Salakhutdinov, "Unsupervised learning of video representations using LSTMs," in *ICML'15: Proceedings of the 32nd International Conference on International Conference on Machine Learning*, 2015, vol. 37, pp. 843–852.
- [17] I. A. Monir, M. W. Fakh, and N. El-Bendary, "Multimodal deep learning model for human handover classification," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 11, no. 2, pp. 974–985, Apr. 2022, doi: 10.11591/eei.v11i2.3690.
- [18] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 1933–1941, doi: 10.1109/CVPR.2016.213.
- [19] O. T.-C. Chen, H. H. Manh, and W.-C. Lai, "Activity recognition of multiple subjects for homecare," in *2018 10th International Conference on Knowledge and Smart Technology (KST)*, Jan. 2018, pp. 242–247, doi: 10.1109/KST.2018.8426164.
- [20] O. T.-C. Chen, C.-H. Tsai, H. H. Manh, and W.-C. Lai, "Activity recognition using a panoramic camera for homecare," in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Aug. 2017, pp. 1–6, doi: 10.1109/AVSS.2017.8078546.
- [21] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, "Efficient kNN classification with different numbers of nearest neighbors," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 5, pp. 1774–1785, May 2018, doi: 10.1109/TNNLS.2017.2673241.
- [22] Z. Lan, M. Lin, X. Li, A. G. Hauptmann, and B. Raj, "Beyond gaussian pyramid: Multi-skip feature stacking for action recognition," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 204–212, doi: 10.1109/CVPR.2015.7298616.
- [23] C. R. de Souza, A. Gaidon, E. Vig, and A. M. López, "Sympathy for the details: Dense trajectories and hybrid classification architectures for action recognition," in *Computer Vision – ECCV 2016*, 2016, pp. 697–716.
- [24] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *2013 IEEE International Conference on Computer Vision*, Dec. 2013, pp. 3551–3558, doi: 10.1109/ICCV.2013.441.
- [25] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 4597–4605, doi: 10.1109/ICCV.2015.522.
- [26] L. Wang, Y. Qiao, and X. Tang, "MoFAP: A motion for action recognition," *International Journal of Computer Vision*, vol. 119, no. 3, pp. 254–271, Sep. 2016, doi: 10.1007/s11263-015-0859-0.
- [27] Y.-G. Jiang, Q. Dai, W. Liu, X. Xue, and C.-W. Ngo, "Human action recognition in unconstrained videos by explicit motion modeling," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3781–3795, Nov. 2015, doi: 10.1109/TIP.2015.2456412.
- [28] D. Li, T. Yao, L.-Y. Duan, T. Mei, and Y. Rui, "Unified spatio-temporal attention networks for action recognition in videos," *IEEE Transactions on Multimedia*, vol. 21, no. 2, pp. 416–428, Feb. 2019, doi: 10.1109/TMM.2018.2862341.




BIOGRAPHIES OF AUTHORS

Manh-Hung Ha    received the M.S. degrees in Information Communication Technology from University of Paris 13, France, and the Ph.D. degree with the Department of Electrical Engineering, National Chung Cheng University, Taiwan in 2014 and 2021, respectively. He was Lecturer with the Faculty of Electrical Engineering, Phenikaa University, Hanoi, Vietnam, from September 2021 to July 2022. Since July 2022, he has been a Lecturer with the Faculty of Application Information Technology, International School, Vietnam National University, Hanoi, Vietnam. His major research interests include multimedia/image/video analytics, computer vision, speech signal processing, and machine learning. He can be contacted at email: hunghm@vnu.edu.vn.






The-Anh Pham    obtained a master's degree in information technology from the University of Limoges, France in 2016. After that, he carried out a Ph.D. thesis in the IRISA/INRIA laboratory from 2016 to 2019 and got a Ph.D. in computer science from the École normale supérieure de Rennes, France, in 2019. He was a postdoctoral researcher at the GSSI Institute, Italy, and the German Aerospace Center (DLR), Germany, in 2020 and 2021 respectively. He is currently a formal methods developer at Prover Technology AB, Sweden. His research theme concerns machine learning, computer vision, the specification, verification of concurrent, and interlocking systems. He can be contacted at the.anh.pham@prover.com.



Dao Thi Thanh    received her M.S. degree in Department of Computer Science and Information Engineering at the National Chung Cheng University, Taiwan in 2017. She is currently a lecture in the Department of Software Engineering at the FPT University (FPT), Hanoi Vietnam. She is also a PhD candidate in the Department of Computer Science and Information Engineering, National Chung Cheng University, Taiwan. Her main research interests include, deep learning, IoT, computer vision, and embedded system design. She can be contacted at email: ThanhDT59@fe.edu.vn.



Van Luan Tran    received the M.A. degree in Electronic Engineering from University of Technology and Education in HCM city, Vietnam, and the Ph.D. degree in electrical engineering from National Chung Cheng University, Chiayi, Taiwan. In 2021, he joined the School of Engineering, Eastern International University, Vietnam, as a lecturer. His research interests include robotics, computer vision, and deep learning. He can be contacted at email: luan.tran@eiu.edu.vn.