IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 11, 2021, accepted November 30, 2021, date of publication December 10, 2021, date of current version December 21, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3134694

Deep Neural Networks Using Residual Fast-Slow Refined Highway and Global Atomic Spatial Attention for Action Recognition and Detection

MANH-HUNG HA[1,2] AND OSCAL TZYH-CHIANG CHEN[1,3], (Senior Member, IEEE)
[1]Department of Electrical Engineering, National Chung Cheng University, Chiayi 621301, Taiwan
[2]Faculty of Electrical and Electronic Engineering, PHENIKAA University, Hanoi 12116, Vietnam
[3]Center for Innovative Research on Aging Society (CIRAS), National Chung Cheng University, Chiayi 621301, Taiwan

Corresponding author: Oscal Tzyh-Chiang Chen (oscal@ee.ccu.edu.tw)

This work was supported in part by the Center for Innovative Research on Aging Society (CIRAS) from the Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE), Taiwan; in part by the Ministry of Science and Technology, Taiwan, under Contract MOST 110-2221-E-194-025 and Contract 109-2622-E-194-004; and in part by PHENIKAA University, Vietnam.

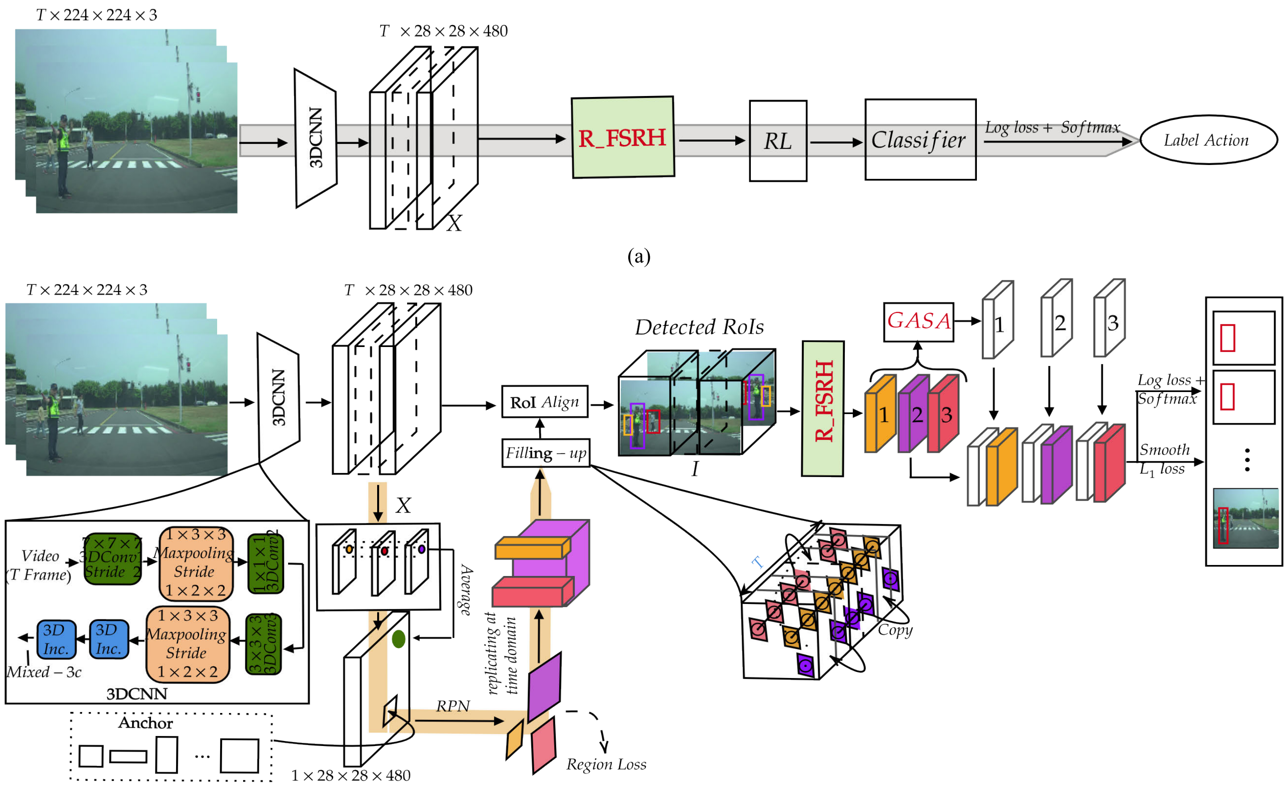
ABSTRACT In this work, we propose two Deep Neural Networks, DNN-1 and DNN-2, based on residual Fast-Slow Refined Highway (FSRH) and Global Atomic Spatial Attention (GASA) to effectively recognize and detect actions. The proposed DNN-1 includes a 3D Convolutional Neural Network (3DCNN), Residual FSRH (R_FSRH), reduction layer, and classification layer for action recognition. In action detection of subject-region extraction and classification, the proposed DNN-2 consists of a 3DCNN, region proposal network, R_FSRH, GASA, and classification-localization layer. The 3DCNN takes the front layer to the ''Mixed-3c'' layer of the pre-trained Inflated 3D (I3D) network as the backbone structure. FSRH is composed of two Refined Highway (RH) units to extract a pair of features from fast and slow actions, where RH has temporal attention from a non-local 3D convolution and an affine transform by a temporal bilinear inception. In R_FSRH, multiple cascaded FSRHs with different residual connections were investigated to determine an effective one. GASA sequentially computes and concatenates the correlation features of an atomic subject and other subjects to effectively discover high-level semantic information. In ablation studies, extensive experiments were conducted to demonstrate the superior performance of the proposed DNN-1 and DNN-2 on five challenging video datasets of JHMDB-21, UCF101-24, Traffic Police (TP), Charades, and AVA. Notably, the proposed DNN-1 shows state-of-the-art performance of 98.6% on UCF101-24 and 98.1% on TP, and DNN-2 exhibits state-of-the-art video-mAP of 27.7% on AVA, and the second best video-mAP of 25.7% on Charades, to the best of our knowledge. Therefore, the DNN-1 and DNN-2 proposed herein can be outstanding context-aware engines for various video understanding applications.

INDEX TERMS 3D CNN, inception network, highway network, residual network, attention mechanism, action recognition, action detection.

I. INTRODUCTION

In context-aware video applications, the complicated interactions among subjects, objects, and environments in a scene may not be easily perceived. In particular, it is still challenging to recognize subject actions that involve multiple subjects' behaviors and interactions. These interactions, which are not always observable in a single frame, require multiple consecutive frames for further comprehension. Recently, deep learning has been successfully demonstrated with impressive performance in video understanding. However, the learning process used to demand large and diverse data in which collection and annotation tasks take a lot of time. Additionally, it is extremely time-consuming to train deep neural networks on massive video datasets. Accordingly, many researchers have dedicated efforts to

The associate editor coordinating the review of this manuscript and approving it for publication was Frederico Guimarães.



(a)

(b)

FIGURE 1. Proposed deep neural networks for action understanding. (a) Actions recognition (b) Action detection.

realize efficient recognition using a limited amount of data [1], and even to employ unsupervised approaches to deal with large-scale detection tasks [2].

Previous studies used Convolutional Neural Networks (CNNs) to handle spatial and temporal features individually. To perceive long-term information in a video, Recurrent Neural Networks (RNNs) are integrated with 2D CNNs to dig out temporal correspondence [3]. In contrast to 2D CNNs, 3D CNNs capture spatial and temporal information simultaneously using 3D spatiotemporal kernels [4], which are beneficial to video perception. Accordingly, 3D CNNs are popularly employed in the field of action recognition, where many improved networks have been derived [5]–[9]. Spatiotemporal 3D CNNs were effectively fulfilled in action recognition, where structural and semantic features were addressed [6], [10]. Alternatively, several pose-based network structures were successfully developed for action classification [7].

Intrinsically, 3D CNNs are computationally complex. There are some available structural modifications aimed at reducing the computational complexity of 3D CNNs. A separable 3D CNN factorized a 3D convolution into 2D and 1D convolutions [11]. The multi-fiber network employed multiple-path 2D convolutions with multiplexers to reduce considerable computation and maintain similar performance [12]. The temporal shift module moved feature maps along the temporal domain, which were processed by 2D CNNs to accommodate temporal information [13]. Their computation and parameter amounts were similar to those of commonly used 2D CNNs.

To ameliorate application performance, some very deep neural networks are quite promising. However, it is challenging to effectively train a very deep neural network. To address training efficiency, the highway mechanism adopted in a network with hundreds of stacked layers was successfully demonstrated using stochastic gradient descent learning [14]. Each highway unit allows extensive flexibility in controlling the transformation and transport of inputs. The flexibility of highway units is likely to contribute to performance improvement, even in a network without a relatively high depth [15]. However, these studies are based on local operations for short-term contextual information extraction and expand the receptive field by stacking multiple units, resulting in inefficiency and difficulty in obtaining long-term contextual information. To overcome this issue, we propose an improvement means associated with feature transform and carry to allow the network to go deeper for preferable performance and fair computational complexity, in which a Refined Highway (RH) unit is designed to yield transformed and carried features.

Generally, subject actions involve short- and long-time movement characteristics. Every frame associated with a short-time action is relatively significant, whereas the onset and offset frames for a long-time action may not be necessary. For instance, falling occurs quickly, whereas jogging may take a very long time. Therefore, fast and slow actions can be partitioned and then identified separately [16], [17]. Additionally, most subject actions involve both short- and long-time behaviors. Behavior features at short and long time need to be treated simultaneously. Accordingly, we devise a Fast-Slow Refined Highway (FSRH) block based on two RH units to extract a pair of features from fast and slow actions. Furthermore, various stacking topologies of multiple FSRH blocks in the Residual FSRH (R_FSRH) module are explored to find out an effective one according to the ideas from the Gated Recurrent Unit (GRU) and residual networks.

For subject detection, Region-based Convolutional Neural Networks (R-CNN) series have been popularly used to extract region proposals by selective searches or Region Proposal Network (RPN) [18]–[20]. Inspired by RPN, we employ a sampling method to localize the subjects who perform actions. In addition, the detector is fulfilled over frames to generate a set of detection results. Subsequently, the Region of Interest Alignment (RoIAlign) scheme is used to adjust the misalignment of all detected persons according to their features [19]. Bounding boxes of detected persons, which are likely the cores of actions, are applied together to find subject interactions as well as individual actions. The same person with multiple interactions with others in a scene leads to multiple action labels. Hence, it is essential for action detection to characterize the interaction context.

It is beneficial to implement subject detection and action classification using scene context [19]–[21]. Intuitively, semantic clues of actions from the scene context can be adopted to ameliorate action identification performance. We incorporate the scene context by correlating the region features of an atomic subject with those of the other subjects in a scene. The features from the other subjects provide the global interaction context useful for the action detection of an atomic subject. Therefore, the Global Atomic Spatial Attention (GASA) module is designed to discover contextual scene descriptors, including high-level semantic information from all discovered subjects. A single-class label per task is often insufficient to describe the content of a video. Even a single frame may contain several prominent persons with multiple labels. Accordingly, the proposed GASA is an effective approach for exploring the interaction relationship between an atomic subject and other subjects in an environment for efficient multi-label action identification.

According to the abovementioned innovative units, blocks, and modules, we implement two tasks: (i) an action recognition task that only estimates the category label for a video clip without localization, as shown in Fig. 1(a), and (ii) an action detection task that not only estimates the category of an action but also localizes an action instance in Fig. 1(b). In action recognition, the proposed Deep Neural Network at type I (DNN-1) includes a 3DCNN, R_FSRH, Reduction Layer (RL), and classification layer. In action detection, the proposed DNN at type II (DNN-2) consists of a 3DCNN, RPN, R_FSRH, GASA, and classification-localization layer. 3DCNN is based on the front layer to the "Mixed-3c" layer of the pre-trained Inflated 3D (I3D) network [20]. In R_FSRH, multiple cascaded FSRH blocks with different residual connections are investigated to determine an effective one according to the ideas from GRU and residual networks in favor of simplicity and ease of training. GASA computes subject-correlated features with high-level semantic information for effective action discernment. In ablation studies, extensive experiments were conducted to demonstrate the outstanding performance of the proposed DNN-1 and DNN-2 on five challenging video datasets of JHMDB-21 [18], [22], UCF101-24 [18], [22], Traffic Police (TP) [23], [24], Charades video classification [25], and AVA spatiotemporal action localization and classification [20].

To clearly illustrate our work, the remainder of this paper is organized as follows. Section II provides a brief review of previous researches related to the proposed DNNs. In Section III, the proposed framework is introduced and explained in detail. Section IV addresses the experimental results for large video datasets of JHMDB-21, UCF101-24, TP, Charades, and AVA. The analyses, comparison, discussion, and visualization of the experimental results are presented. Finally, the work is concluded in Section V.

II. RELATED PREVIOUS WORK

A. ACTION UNDERSTANDING VIA DNNS

3D CNNs that explore the features from both spatial and temporal domains simultaneously have been increasingly implemented in video analysis tasks. 3D CNN aims to perform spatial and temporal feature learning together [26], and RNNs are used to capture temporal context and handle variable-length video sequences [27]. However, 3D CNN architectures have many more parameters than 2D CNNs, making them computationally expensive. Inspired by residual networks, skip connections over layers can be applied to 3D CNNs to overcome the problem of vanishing gradients [28]. Another effective approach is to localize subjects on a temporal scale regarding their actions. Many architectures, such as fast temporal action proposals and region convolutional 3D network [29], have also been developed.

B. ATTENTION MECHANISMS

Numerous specialized attention mechanisms have demonstrated promising performance in computer vision tasks [30]. There is also a lot of interest in combining CNNs with various forms of self-attention for image classification [31], object detection [32], and video action recognition [9], [33]. Attention is an effective way to capture short- and long-range dependencies that are commonly used in CNNs to boost the performance of image classification and scene segmentation [34]. Spatial attention emphasizes the spatial relationship among features, while channel attention enhances the most meaningful channels and weakens the others. The approach using the channel-wise attention mechanism [35] was advantageous for increasing the CNN's performance at a low computational cost. On the other hand, non-local block operation considers spatiotemporal information to learn the dependencies of features across frames for video classification tasks [36]. Such non-local block operation can be viewed as a self-attention strategy. For action detection, a hierarchical dilated attention layer was developed in which attention weights were allocated to local frames at multi-temporal scales to improve the quality of local feature representation across time [37].

The proposed DNNs are devised by referring to slow-fast attention [17], highway network, temporal bilinear, and non-local attention of transformer [21], [38]. Because the relation between an atomic subject and its surrounding subjects is employed, the attention from the subject interaction context highlights the relevant parts in a feature map. In particular, the proposed attention mechanism, working as a conditioned feature provision, can yield context information effectively.

C. BOUNDING BOX DETECTORS FOR ACTIONS DETECTION

The task of action detection aims to discern the actions of interest that are present in a video, and to localize them in both spatial and temporal domains. Intuitively, action detection consists of two major steps: (i) a subject related to an action is found by a region scheme or dense sampling via anchors, and (ii) the discovered subject is localized with refinement and identified. In the literature, R-CNN series extracted region proposals using the selective search or RPN at the first stage and then classified subjects from these potential regions at the second stage [18]. Owing to the success of the R-CNN series, most of the studies detected subjects in each frame first and then adequately linked these bounding boxes as region tubes [18]. Although these networks have achieved promising results, the temporal properties of videos are not explicit or fully exploited for detection. To better leverage temporal cues, several recent studies have been developed to carry out action detection using video clips. For instance, the action tubelet approach takes a short sequence of frames as an input and yields regressed tubelets, which are then linked by a tubelet linking scheme to construct action tubes [39]. Gu *et al*. further demonstrated the importance of temporal information by using a long video clip and taking advantage of I3D pre-trained on a large-scale video dataset [20]. Rather than linking detection results based on a frame or segment level, some schemes connect the action tube proposals from multiple video segments before classification [40]. Grid CNN (G-CNN) trained a repressor to iteratively move a grid of bounding boxes toward objects [41]. Cascade R-CNN presented a cascade framework for highly promising object detection, where a sequence of R-CNN detectors was trained with an increased threshold of Intersection over Union (IoU) to iteratively suppress false positives [42]. In context recognition, a two-pathway slow-fast structure was introduced at low and high temporal resolutions to flexibly and effectively characterize videos [16]. The approach took a long-term feature bank with supportive information extracted over the entire video clip to comprehend subject actions well [43]. In general, the task of action detection for each subject in a video clip is more complicated than that of action recognition for a video clip with subject(s). It requires accurate localization and precise categorization of the subjects at a time interval. In contrast to previous work, this study proposes an effective framework for spatiotemporal action detection.

III. PROPOSED DNNs

To realize action recognition, the proposed DNN-1 has a 3DCNN, R_FSRH, RL, and classification layer, as shown in Fig. 1(a). The 3DCNN based on the 1st layer to the ''Mixed-3c'' layer of the pre-trained I3D network processes an input of a video segment with T frames. The output from 3DCNN forms a 3D feature map X with a dimension of (T, H, W, D), where H and W represent the height and width, respectively, and D denotes the number of channels. To obtain the long-term dependency, X feeds into the proposed R_FSRH module for context extraction. By stacking multiple FSRH blocks with residual connections, R_FSRH can extract dense long-term spatiotemporal relation information of which dimension is reduced by RL. The classification layer is a two-layer feedforward neural network.

In action detection, the proposed DNN-2 in Fig. 1(b) includes a 3DCNN, RPN, R_FSRH, GASA, and classification-localization layer. Here, 3DCNN, whose structure is the same as that used for action recognition, generates features for RPN to look for subjects present in a scene. The first, middle, and last frames of the feature map in the temporal domain are selected, and element-wise averaged at the spatial domain to yield the averaged feature frame that passes through RPN. The region proposals from RPN are processed to produce a region feature tube using the proposed Bounding-box Filling-up (BF) procedure and the RoIAlign scheme. Subsequently, R_FSRH is employed to unveil the short- and long-term context features from each region feature tube related to a subject, and GASA is used to enrich the relationship features between an atomic subject and its surrounding subjects. Meanwhile, the region feature tube is regressed to form a 4D offset vector to adjust the region proposal into a tight bounding box around a person. Finally, the action classification-regression layer adopts two two-layer fully connected networks in which there are C action classes and background (total $C + 1$).

A. PROPOSED R_FSRH

R_FSRH inspired by the GRU, highway networks, and residual networks consisting of multiple stacked FSRHs with residual connections are presented as follows:

1) PROPOSED FAST-SLOW REFINED HIGHWAY BLOCK

a: RH CONCEPT

The transform and carry gates in traditional highway networks adopt simple nonlinear operations, such as a sigmoid function. Fig. 2 shows the conceptual block diagram of the proposed RH unit that extends highway networks with two

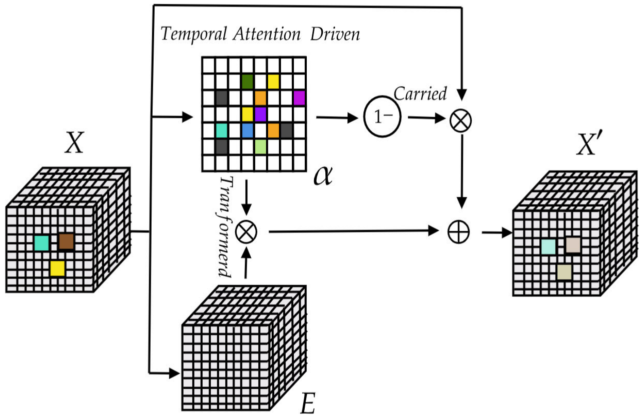


FIGURE 2. Conceptual block diagram of the RH unit for 3D attention.

new traits: (1) temporal attention driven parameters, $\boldsymbol{\alpha}$ providing weighting ratios of transformed and carried outputs, are produced by a non-local 3D convolution, and (2) the affine transform, E, is computed by operations such as Temporal Convolution (TC), Temporal Bilinear (TB), and Temporal Bilinear Inception (TBInc). Temporal attention driven parameters, $\boldsymbol{\alpha}$, are normalized by an activation function of SoftMax to obtain values in the range of 0 to 1, where each parameter can be dynamically adjusted according to temporally shifted features. The output from E takes a matrix multiplication with attention weights, $\boldsymbol{\alpha}$, to yield the transformed one while a given input, X, performs a matrix multiplication with 1-$\boldsymbol{\alpha}$ to yield the carried one. Afterwards, the transformed and carried ones are element-wise added together to produce the output X'. In particular, if $\boldsymbol{\alpha}$ is close to zero in the transform state, then 1- $\boldsymbol{\alpha}$ for the carry state approaches one. Such a case directly passes an input X to the output in which RH works like a write gate of GRU. Intuitively, the RH unit behaves like a variant of the highway network, GRU, and residual network. In the following, TC, TB, and TBInc used in E are further presented.

b: TEMPORAL CONVOLUTION (TC)

T feature frames can be represented by $X^1, X^2, \ldots, X^T$ at $X^i \in \mathbb{R}^D$. In this work, these features are aggregated in the temporal domain. The output signals are defined as $Y^1, Y^2, \ldots, Y^{T'}$ where $Y^i \in \mathbb{R}^D$, and T' is the temporal output dimension. In practice, each output feature corresponds to several consecutive input feature frames. Similar to [44], the temporal convolution is extended from the spatial convolution operator. It performs a learnable transformation on the input feature frames as follows:

$$y_C = \sum_{j=1-\left\lfloor \frac{k+1}{2} \right\rfloor}^{\left\lfloor \frac{k}{2} \right\rfloor} W_c^j X^{i+j} \tag{1}$$

where W_c^j is the weight matrix for the c^{th} output neuron, and k is the number of input feature frames. Based on convolutions at $(3 \times 1 \times 1)$, TC obtains the transformation within T feature frames.

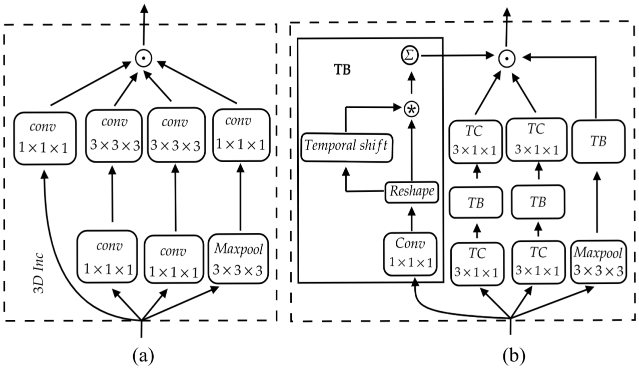


FIGURE 3. Inception units. (a) 3D Inception (Inc.). (b) Proposed Temporal Bilinear Inception (TBInc).

c: TEMPORAL BILINEAR INCEPTION (TBInc)

Based on the 3D inception in Fig. 3(a), the learning of spatial and temporal features needs to be fairly addressed to balance spatial and temporal clues for good performance. Because the temporal information stored in the shifted channels may be lost owing to a large number of channels, TB operations between two TCs in a TBInc, as depicted in Fig. 3(b), are considered. In TB, 3D convolution first calculates the transformation. Second, a temporal shift operator based on a time index is implemented using translation operations. Element-wise multiplication and summation are performed to obtain the final result of TB for all spatial and temporal elements. TBs are embedded in four branches of an inception unit. In each TB, a temporal shift can address degraded spatial feature learning. Fig. 3(b) displays the concatenation topology of four branches including four TBs in which two are inserted between two pairs of two TC operators at $3 \times 1 \times 1$, one is appended after the operator of maxpool at $3 \times 3 \times 3$, and one is appended in the branch without the other operation.

d: TEMPORAL ATTENTION DRIVEN (TAD)

The operations of TAD in Fig. 4 are executed as follows. $\boldsymbol{X}$ denotes the 3D feature map with a dimension of $T \times k \times k \times D$, which are transformed to $(T \times D) \times (k \times k)$ feature tensors at three different 3D convolutional layers with kernels at a size of $1 \times 1 \times 1$ to generate feature maps of P, Q and U where $P, Q, U \in \mathbb{R}^{(T \times D) \times (k \times k)}$. $k \times k$ is the dimension of a spatial feature map. P and Q are used to calculate the attention weights of $\boldsymbol{\alpha} \in \mathbb{R}^{(T \times D) \times (T \times D)}$ by matrix multiplications. $\alpha_{i,j}$ is the relationship intensity of the i^{th} location to the j^{th} region at $i, j \in (1, \ldots, T \times D)$, represented by the following equations:

$$\alpha_{i,j} = \frac{\exp\left(V_{i,j}\right)}{\sum_{i=1}^{T \times D} \exp\left(V_{i,j}\right)} \tag{2}$$

$$V = PQ^T \tag{3}$$

Subsequently, $\boldsymbol{\alpha}$ is convoluted with U to obtain the output that goes through a 3D convolution at a $1 \times 1 \times 1$ kernel to yield $\boldsymbol{m} \in \mathbb{R}^{T \times k \times k \times D}$.

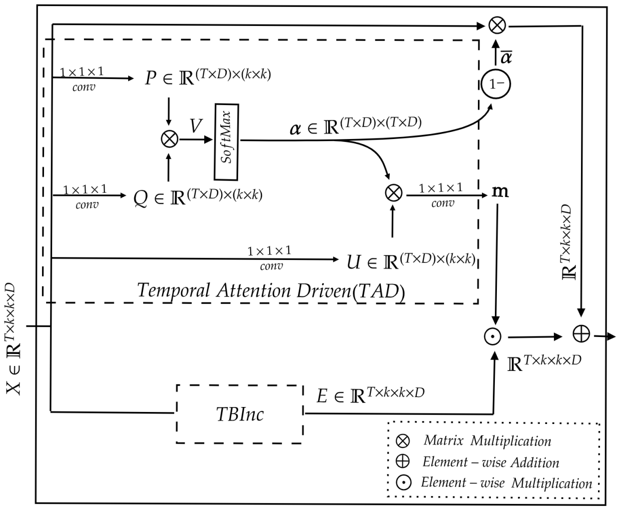

$X \in \mathbb{R}^{T\times k\times k\times D}$
1×1×1 conv
$P \in \mathbb{R}^{(T\times D)\times(k\times k)}$
V
SoftMax
$\alpha \in \mathbb{R}^{(T\times D)\times(T\times D)}$
$\bar{\alpha}$
1−
1×1×1 conv
$Q \in \mathbb{R}^{(T\times D)\times(k\times k)}$
1×1×1 conv
m
1×1×1 conv
$U \in \mathbb{R}^{(T\times D)\times(k\times k)}$
Temporal Attention Driven(TAD)
$\mathbb{R}^{T\times k\times k\times D}$
$\mathbb{R}^{T\times k\times k\times D}$
$X' \in \mathbb{R}^{T\times k\times k\times D}$
TBInc
$E \in \mathbb{R}^{T\times k\times k\times D}$
⊗ Matrix Multiplication
⊕ Element – wise Addition
⊙ Element – wise Multiplication

FIGURE 4. Fast refined highway.

e: FAST RH (FRH) UNIT

As shown in Figs. 3 and 4, $\boldsymbol{X}$ drives TBInc, which considers the temporal pairwise feature interactions across adjacent feature-level frames to increase the capability of characterizing spatiotemporal dependencies. Bilinear with mixture inception is capable of capturing temporal interactions. The output of TBInc is the cascaded four outputs of inception operations at four branches to generate a feature tensor with a dimension of $T \times k \times k \times D$. Therefore, TBInc can be easily implemented and incorporated into the widely used neural network architectures.

By looking at conventional highway networks, their approaches exhibited how an output was produced by transforming and carrying an input. We refer to the transform and carry gates by weight multiplications of $\boldsymbol{m}$ and $1 - \boldsymbol{\alpha}$, respectively, in Fig. 4. The computation of the RH unit is defined as

$$X' = \boldsymbol{m} \odot E(X) \oplus (1 - \boldsymbol{\alpha}) \otimes X \tag{4}$$

where $\odot$, $\oplus$ and $\otimes$ denote element-wise multiplication, element-wise addition and matrix multiplication, respectively. With $\boldsymbol{\alpha}$ ranging from 0 to 1, 1-$\boldsymbol{\alpha}$ is defined as $\bar{\alpha}$. Therefore, the output, X', is the weighted summation of the transformed and bypassed inputs according to the attention weights of $\boldsymbol{m}$ and $\bar{\boldsymbol{\alpha}}$, respectively, where $\boldsymbol{m}$ represents significant context parts of $\boldsymbol{\alpha} \otimes U$ for the transform gate.

f: FSRH BLOCK

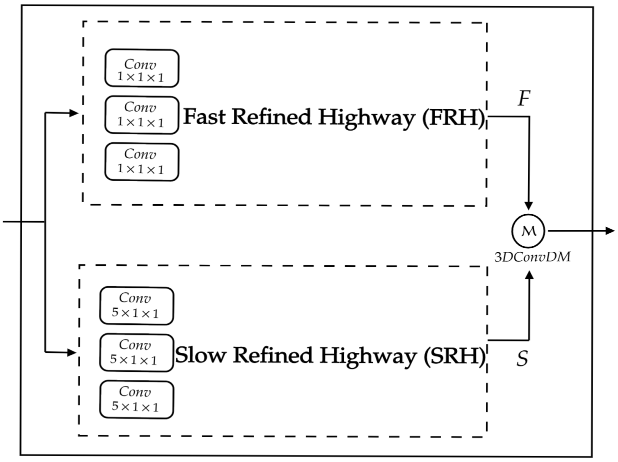


FIGURE 5. Fast-slow refined highway.

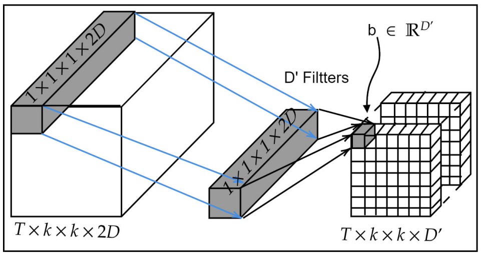


FIGURE 6. 3DConvDM.

In parallel to a fast action, a slow action is also considered where convolution filters in the FRH unit are replaced by $5 \times 1 \times 1$. In Fig. 5, FSRH extracts two types of features that distinguish fast and slow actions, where the kernel sizes of the convolution operations are $1 \times 1 \times 1$ and $5 \times 1 \times 1$ adopted in the Fast Refined Highway (FRH) and the Slow Refined Highway (SRH) units, respectively. In order to increase the relevant feature extraction, the outputs from FRH and SRH are concatenated and then correlated by 3D Convolutional Depth Mapping (3DConvDM). The 3DConDM in Fig. 6 processes two output feature maps of F and S stacked at the same spatial locations across the feature channels to become $[F, S]$ at a dimension of $T \times k \times k \times 2D$, which are element-wise convoluted with the 3D filters of $R^{1\times1\times1\times2D}$, and then added with biases $b \in \mathbb{R}^{D'}$. Here, the filter of a 3D convolution $(1 \times 1 \times 1)$ to obtain the convolution feature map is used to reduce the dimension by a factor of two and is able to model weighted combinations of two feature maps of F and S at the same spatial location. There are D' of 3D filters at a size of $1 \times 1 \times 1 \times 2D$ to characterize the correlation of the stacked feature maps where these 3D filter matrices are slid in the temporal, spatial-horizontal, and spatial-vertical directions of $T \times k \times k$, to yield outputs after convolutions. The results from D' filters are stacked together to form an output at a size of $T \times k \times k \times D'$ where the stride and padding are equal to one. In this work, the number of 3D filters, D', is equal to D. 3DConvDM correlates two input feature maps to yield an output feature map with meaningful contextual information and selectively aggregated context.

2) RESIDUAL FAST SLOW REFINED HIGHWAY (R_FSRH) MODULE

FSRH generates a feature map with refined spatiotemporal information that may not be powerful enough for accurate action classification. To obtain richer and denser context information, multiple FSRHs with a proper stacking topology

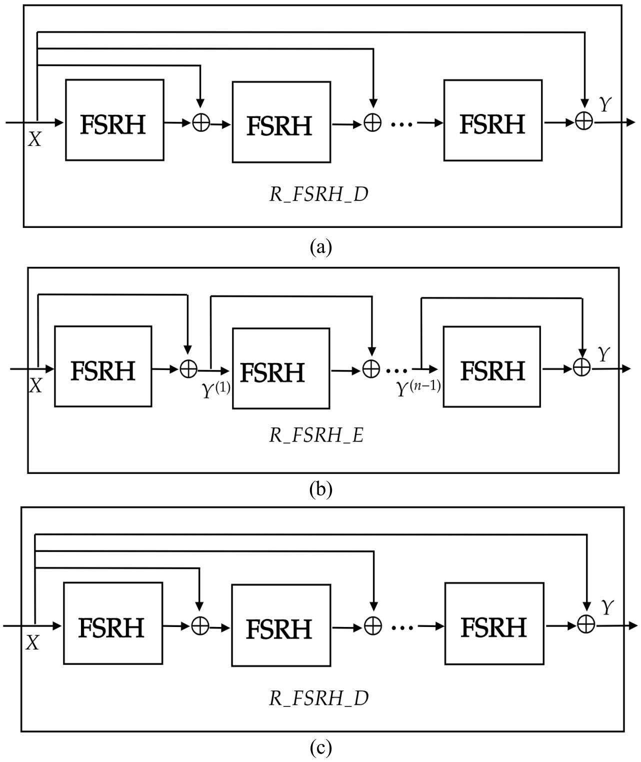


FIGURE 7. R_FSRHs at three types. (a) R_FSRH_S. (b) R_FSRH_E. (c) R_FSRH_D.

are investigated, where these FSRH blocks share the same parameters to avoid too many extra parameters. Therefore, R_FSRH modules having one or multiple cascaded FSRH blocks with residual connections are constructed and analyzed at different cascade depths. R_FSRHs at three types of residual connections, as depicted in Fig. 7, were studied. The first structure is R_FSRH with a Single residual connection (R_FSRH_S). The second is R_FSRH with a residual connection at Each FSRH (R_FSRH_E). The third one adopts residual connections from an initial input to the outputs of all FSRHs, named R_FSRH with Dense connections (R_FSRH_D). Since the appearance information in an input feature map after convolutions may be lost to some extent, an input feature map is added to the output(s) of FSRH(s) to effectively restore the appearance information. Accordingly, the next layer can attain both appearance and context information.

To explore the scaling factor associated with the outputs of FSRHs, three types of R_FSRH are analyzed, where R represents the operation of FSRH. We multiply the output of R_FSRH using a trainable scalar parameter and add the result back to the input feature map as a residual block.

(1) R_FSRH_S in Fig. 7(a) can be interpreted by

$$Y = \gamma R\left(\gamma R\left(\gamma R\left(\ldots\right)\right)\right) + X \tag{5}$$

in which γ is a learnable weight.

(2) In R_FSRH_E, as depicted in Fig. 7(b), only the first FSRH block has an input with the original appearance information. The operations of R_FSRH_E are formulated as follows:

$$Y = \gamma R\left(\gamma R\left(\ldots\left(\gamma R\left(X\right)+X\right)\ldots\right)+Y^{(n-2)}\right)+Y^{(\mathrm{n}-1)} \tag{6}$$

where the input of each FSRH block is added to its output to form a residual connection.

(3) Fig. 7(c) displays R_FSRH_D of which operations are

$$Y = \gamma R\left(\gamma R\left(\gamma R\left(\ldots\right) + X\right) + X\right) + X \tag{7}$$

where all the FSRH blocks have inputs from the original appearance information. These three structures are compared in the next section.

All FSRH blocks in R_FSRHs share weights so that they can make good use of the learned relationship information as well as reduce considerable extra parameters. Because the input and output dimensions of the FSRH block are the same, the proposed FSRH block can be inserted into any DNN to provide an inception layer at any stage. Such an arrangement can meaningfully address the spatiotemporal context features for performance enhancement.

B. RPN, BF AND GASA

The 3DCNN uses a video segment as an input and generates a spatiotemporal feature map. The first, middle, and last frames extracted from the feature map in the temporal domain are averaged to yield the feature frame that is processed by the RPN to produce region proposals associated with subjects. This task is based on faster R-CNN for subject detection [19], [45]. A region proposal provides a person's bounding box. The Bounding-box Filling-up (BF) procedure is proposed to generate a region feature tube that is scaled to a specific size by the RoIAlign scheme. The region feature tube goes through R_FSRH_D and GASA to establish semantic information among each atomic subject and the other subjects related to activities. The resulting feature map then inputs two fully connected layers for classification and bounding box regression, in which the bounding box is regressed by a 4D offset vector to achieve a tight box around a person. The tight bounding box is then classified into multiple labels of C actions and background.

1) COMPUTATION OF BOUNDING BOX VIA RPN

The modified tube proposal network, I, inspired by RPN in faster R-CNN, is developed. The representative feature tube corresponding to the RPN proposals is extracted from the feature map using BF and RoIAlign. Given the feature map from 3DCNN, the classification and regression labels are defined with respect to the tube anchors. Tube anchors that are similar to the bounding box anchors used in faster R-CNN are employed in the temporal domain. This indicates that the tube anchors are implemented by duplicating the boxes to form a tube. The RPN generates multiple potential person bounding boxes, along with objectness scores based on class probabilities and IoU values. The box with the highest objectness score is selected and further regressed into a tight bounding box. The resulting features across time are stacked to yield a spatiotemporal feature tube that passes through the R_FSRH module, as depicted in Fig. 1.

2) BOUNDING-BOX FILLING-UP PROCEDURE

The bounding boxes related to the same subject in the neighboring frames may sometimes be missing. Accordingly, the linking and interpolation of bounding boxes are designed to yield a reliable region feature tube. The proposed BF procedure for generating a region feature tube is as follows. Initially, the BF procedure starts to seek the bounding boxes from the middle of T frames in the temporal domain and correlates them to those found at the left and right neighboring frames. If there are some mismatched bounding boxes, the missing boxes at specific frames are reproduced by copying the boxes from their close neighboring frames. The above operations are repeated until two boundary frames are reached. This approach makes T frames with a consistent number of bounding boxes. Since these boxes may have different sizes, the RoIAlign scheme is used to link and normalize the corresponding bounding boxes at the same size.

3) GASA MODULE

The tasks of action classification and localization have substantially different objectives and require different types of information. Accurate action classification demands applicable context features in both the spatial and temporal domains. Robust localization regression requires precise spatial cues at the frame level. In this work, the output from R_FSRH_D for each atomic subject addresses the GASA module to yield a spatiotemporal correlation feature of an atomic subject and the other subjects. As shown in Fig. 1, the correlation feature from GASA is further concatenated with the feature map of the corresponding atomic subject in the feature frame domain for the classification-regression layer. This concatenated feature not only captures the correlative context among subjects but also includes the local details of an atomic subject for ameliorating action classification and localization regression.

The region feature tube carries information about the subject appearing in a scene. Some actions involve multiple subjects. The subject properties of multiple regions feature tubes, which concentrate on the actions of people, are integrated and processed by GASA, as depicted in Fig. 8. The GASA module obtains the reference information from the first subject and another by using the Atomic Spatial Attention (ASA) block, $ASA\left(A_t^1, S_2\right)$. Each ASA accepts inputs of A_t^i and S_{i+1} where A_t^i is the output from $ASA(A_t^{i-1}, S_i)$ and S_i is the feature tube of another subject. Here, S_i is randomly selected from the list of the available region feature tubes. Each GASA takes care of one atomic subject that is sequentially chosen from the discovered region feature tubes to yield a spatiotemporal subject-related feature.

In Fig. 8, two region feature tubes in ASA-1 are projected into three new feature tensors, P', Q' and U'. The attention map $\boldsymbol{\beta} \in \mathbb{R}^{(k' \times k') \times (k' \times k')}$ is generated by the computations of $P'^T Q'$ and Softmax. Then, $\boldsymbol{\beta}$ is multiplied with U' by a matrix operation to yield the result that goes through a 3D convolution at a 1 × 1 × 1 kernel to construct $\boldsymbol{m}' \in \mathbb{R}^{(T' \times D') \times (k' \times k')}$. Finally, the output of ASA-1 comes from the element-wise summation of A_t^1 and $\boldsymbol{m}'$. The same computation flow is applied to ASA-2, ..., and ASA-n. In particular, we concatenate outputs from n ASAs that contain the global subject correlation from an atomic subject point of view.

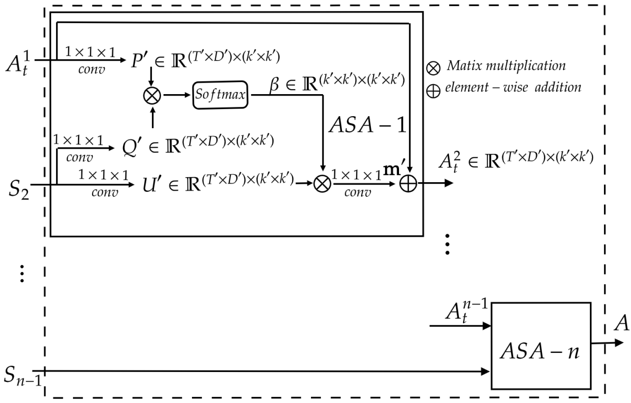

A_t^1
$1\times1\times1$ conv
$P' \in \mathbb{R}^{(T'\times D')\times(k'\times k')}$
Softmax
$\beta \in \mathbb{R}^{(k'\times k')\times(k'\times k')}$
$ASA-1$
$1\times1\times1$ conv
$Q' \in \mathbb{R}^{(T'\times D')\times(k'\times k')}$
S_2
$1\times1\times1$ conv
$U' \in \mathbb{R}^{(T'\times D')\times(k'\times k')}$
$1\times1\times1$ conv
$\mathbf{m}'$
⊗ Matix multiplication
⊕ element – wise addition
$A_t^2 \in \mathbb{R}^{(T'\times D')\times(k'\times k')}$
A_t^{n-1}
$ASA-n$
A_t^n
S_{n-1}

FIGURE 8. GASA module.

C. LOSS FUNCTION OF END-TO-END LEARNING

The RPN extracts candidate bounding boxes to generate region feature tubes. We aggregate and correlate features from a set of regions feature tubes using R_FSRH_D and GASAs. Finally, multiple loss functions are used for detecting subject actions and regressing the bounding box coordinates according to the subject classes and ground-truth bounding boxes, respectively. The regression target is to perceive the subject. The total loss $\mathcal{L}$ of the proposed network is

$$\mathcal{L} = \mathcal{L}_{RPN} + \mathcal{L}_{cls} + \mathcal{L}_{reg} \tag{8}$$

where $\mathcal{L}_{RPN}$, $\mathcal{L}_{cls}$, and $\mathcal{L}_{reg}$ denote the loss of RPN, classification loss for identifying subject actions, and the bounding box regression loss, respectively.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, the proposed DNNs are evaluated and compared with the conventional ones in the tasks of action recognition and detection on several popularly-used video datasets. Additionally, extensive ablation experiments were conducted, analyzed, and discussed.

A. DATASETS

1) DATASETS FOR ACTION RECOGNITION

JHMDB-21 consists of 928 short videos with 21 action categories associated with daily life [18], [22]. Each video was well trimmed and had a single action instance across all frames. Our experiments were performed during the first split. UCF101-24 is a dataset related to subject actions in which there are 3,207 videos with 24 action classes [18], [22].

All experiments using UCF101-24 were fulfilled at the first split as well. In the dataset of Traffic Police (TP) [23], [24], there are 21 video clips, including nine actions associated with how to direct vehicles. The frame resolution and rate of these video clips are 1, 080 × 1.080 pixels and 15 Hz, respectively.

2) DATASETS FOR ACTION DETECTION

The proposed DNN-2 was tested on four challenging datasets: JHMDB-21, UCF101-24, Charades, and AVA, where precise person boxes were adopted in accordance with human silhouettes as well as the ground truths of labeled actions. JHMDB-21 and UCF101-24 annotations refer to the ground truths provided by [18], [22], [45], in which the fair bounding boxes of subjects were offered. In each video, there may be multiple action instances that have the same class label but different spatial and temporal boundaries. Charades [25] and AVA [20] contain the activities of multiple persons in scenes. Each video clip was severely untrimmed and had numerous action labels. These phenomena make both datasets much more challenging. The Charades dataset has 157 classes in 9,848 training and 1,800 validation videos, each of which has multiple labels regarding actions at an average interval of 30 seconds. In each video clip, a person can perform one or more actions. The AVA v2.1 dataset contains 211 K training and 57 K validation video samples, each of which has a 15-minute clip extracted from a movie. Each person in a frame has a bounding box and multiple labels. The task in AVA is spatiotemporal action localization and discernment: each person appearing in a test video is detected in a frame, and the multi-label actions of the detected person are perceived as well. Notably, each bounding box may be associated with multiple action categories, making the dataset more challenging. There exist 60 action classes that are partitioned into three groups of 13, 32, and 15 classes related to person poses, object interactions, and person interactions, respectively. According to the AVA v2.1 evaluation process, the mean Average Precision (mAP) across 60 classes was calculated.

B. IMPLEMENTATION DETAILS

In our experiments, a video clip was partitioned into many video segments, each of which had 10 frames, i.e. $T = 10$. All frames in a video segment were resized to 224 × 224 pixels. We trained our models using Adam [21], and a batch size of 8. The initial learning rate was 5×10^{-5} with a step decay after 20 and 30 epochs at a decay rate of 0.1. In this work, the experiments associated with model training and testing were performed on a 16-core Intel Core i7 computer accompanied with two Graphic Processing Units (GPU) of Nvidia Titan X 1080/12 GB at 64GB RAM, where the CPU system is a 64-bit Ubuntu 18.04 and GPUs are supported by the Anaconda distribution for Python programming.

1) TRAINING DETAILS

Based on the transferred parameters of the I3D network [16], [46], 3DCNNs were retrained for our tasks. By using the RoIAlign scheme, all of the region proposals from the feature map are represented by the fixed-shape feature vectors that pass through the fully connected layers to become the region feature vectors. A proposal is assigned to a ground-truth box based on an IoU criterion greater than 0.5. Proposals with an IoU between 0.1 and 0.5 are treated as negatives whereas proposals with an IoU below 0.1 are ignored. The networks are trained using sigmoid cross-entropy for multi-label classification loss in all proposals at mini-batch, where the standard smooth L1 loss for bounding box regression is employed by localizing a positive proposal. To remove duplicates, a category based on the Non-Maximum Suppression (NMS) scheme was applied [19]. The final output is a list of all predicted subjects with probabilities higher than the IoU threshold. The labels of a ground-truth box are assigned to a predicted box if these two boxes have an overlap of IoU ≥ 0.5. Because a specific person may have multiple actions at a short interval, the output result has a dimension of $G^{*}(C+1)$, where G indicates the number of predicted subjects, C denotes the number of action classes, and an additional class is the background. In practice, the detection outputs at each step are gathered and used to select the positive and negative samples for training in terms of IoU. We accumulate the losses of all steps and back-propagate the entire network to update their weights at the same time.

2) METRICS

In action recognition and detection, the metrics are the average accuracy and video-level mAP at specific IoU thresholds, respectively. In this work, an IoU threshold of 0.5 is adopted to assess the detection performance. Additionally, the computational complexities of FLOPs for network inferences were provided. To avoid unnecessary calculations in the network model, we freeze the model before measuring FLOPs and then sum the operations of all convolution layers based on the numbers of output feature maps, kernel sizes, input channels, and output channels. In particular, the actual inference time is computed by a function of time() at 10 224 × 224 × 3 image frames.

C. ABLATION STUDIES

The proposed DNNs are investigated for action recognition and detection. First, the performance of R_FSRH at different structures in DNN-1 was explored on three datasets. Second, DNN-2 with variant simplified configurations based on FSRH, R_FSRH, and GASA were analyzed and compared. In particular, the proposed R_FSRH module is generic and can be applied to other network topologies for performance improvement.

1) ACTION RECOGNITION ON JHMDB-21, UCF101-24 AND TP

Table 1 lists the accuracy performance of the proposed DNN-1 with and without R_FSRH_D at $\gamma = 1$, as shown in Fig. 7(c), on three datasets where the number, τ, of FSRH blocks ranges from 1 to 5. Compared to the baseline of

TABLE 1. **Accuracies of DNN-1 with and without R_FSRH_D on JHMDB-21, UCF101-24 and TP.**

Ablation of DNN-1	JHMDB -21	UCF101-24		TP
		Accuracies	No. of parameters (Million)	
3DCNN	84.03	96.24	13.78	95.32
3DCNN+ R_FSRH_D (τ=1)	85.91	98.28	17.23	97.93
3DCNN+ R_FSRH_D (τ=2)	86.01	98.47	19.27	98.06
3DCNN+ R_FSRH_D (τ=3)	**86.03**	**98.56**	21.31	**98.13**
3DCNN+ R_FSRH_D (τ=4)	85.87	98.20	23.35	97.77
3DCNN+ R_FSRH_D (τ=5)	85.86	98.08	25.39	97.75

3DCNN, the proposed DNN-1 with R_FSRH_D at $\tau = 1$ increased the accuracies by 1.88%, 2.04%, and 2.61% on JHMDB-21, UCF101-24, and TP, respectively, clearly revealing the contribution of R_FSRH_D. Furthermore, the performance at τ from 1 to 3 is ameliorated from 0.12% to 0.28% on three datasets, demonstrating the effectiveness of dense contextual information. Meanwhile, the number of FSRH blocks increased from three to four, and slightly reduced the performance by 0.16%, 0.36%, and 0.36% on JHMDB-21, UCF101-24, and TP, respectively. Additionally, the number of parameters and the inference time increased with τ. Specifically, when τ is equal to 5, the number of parameters is a maximum of 25.39M but the accuracy decreases compared to that at $\tau = 3$. Therefore, R_FSRH at $\tau = 3$ has a relatively good spatiotemporal interpretation capability, where the network achieves the highest accuracy for the three datasets. The proposed DNN-1 with the R_FSRH_D module at three FSRH blocks achieves the best performance. This effect reveals that the increasing number of FSRH blocks in R_FSRH_D likely tends to diminish the importance of the supportive context. To focus on the performance and resource usage, R_FSRH using three FSRH blocks was adopted in the following experiments.

In the proposed DNN-1, the R_FSRH modules with three FSRH blocks at different residual connections in Fig. 7 were investigated. The experimental results are listed in Table 2. The weight γ in Eqs. (5) to (7) is learnable so that the ratio of inputs to outputs from FSRHs can be adjusted automatically. According to the simulation results, the accuracies varied at different initial values of γ. The experimental settings in these three structures are the same, where γ is set to nine different initial values ranging from none to 1.0. When γ is none, there is no γ in Eqs. (5)–(7), indicating R_FSRH without a learnable weight. The experimental results show that the top-1 accuracies of the proposed R_FSRH_D module at γ =1 are 86.03% and 98.56% for JHMDB-21 and UFC101-24, respectively. This indicates that the initial contribution from FSRHs is the same as that from the residual paths, which is beneficial for good convergent performance. Therefore, we adopted the initial value of γ as 1 for the other experiments. Compared with R_FSRH_S and R_FSRH_E, the structure of R_FSRH_D makes good use of the dense correlation information in capturing long-range dependencies using three FSRH blocks. Because the accuracy of R_FSRH_D is higher than that of the others, the proposed DNNs adopt R_FSRH_D.

TABLE 2. **Accuracies of DNN-1 using three types of R_FSRHs at different γ on JHMDB-21 and UCF101-24.**

Initial γ	R_FSRH_S (JHMDB-21| UCF101-24)	R_FSRH_E (JHMDB-21| UCF101-24)	R_FSRH_D (JHMDB-21| UCF101-24)
none	85.00|98.30	84.90|98.37	85.01|98.45
0.001	85.00|98.30	84.91|98.38	85.02|98.48
0.05	85.00|98.31	84.91|98.39	85.02|98.49
0.1	85.10|98.39	84.92|98.40	85.11|98.42
0.25	86.00|98.43	85.99|98.41	86.01|98.45
0.5	86.00|98.45	86.00|98.43	86.00|98.45
0.75	86.01|98.48	86.00|98.44	86.01|98.49
0.9	86.02|98.49	86.01|98.46	86.02|98.50
1.0	86.02|98.53	86.01|98.53	**86.03|98.56**

2) ACTION DETECTION ON JHMDB-21, UCF101-24, AND CHARADES

To verify the performance of the proposed DNN-2 in the action detection task, we conducted extensive ablation experiments on the datasets of JHMDB-21, UCF101-24, and Charades.

a: ANALYSES OF FRHs WITH VARIANT CONFIGURATIONS

In the experiments of action detection using DNN-2 at $\tau = 1$, the performance of FRHs with variant configurations was analyzed and compared, where FSRH was simplified to FRH for analysis purposes. Region candidates associated with subjects were selected when IoU was greater than or equal to 0.5. Additionally, Giga FLOPs (GFLOPs) and Number of Parameters (NPs) for each configuration are listed in Table 3. As shown in Fig. 4, TAD and TBInc of the FRH unit provide the transform and carry features revealing how the output is produced by transforming and carrying an input, respectively. To explore variant configurations, an FRH unit can be implemented by a single or pair component, which includes one of TAD and Attention Mechanism (AM) [5], and/or one of TBInc, Inc. and TB, as depicted in Fig. 3. From the list of outcomes associated with FRHs having only a single unit in Table 3, FRH using TAD achieves the highest video-mAP values on the three datasets. Particularly, the GFLOPs and NPs of TAD are less than those of TBInc

TABLE 3. Performance of DNN-2 using variant FRHs at τ =1.

Carrier		Transformer					JHMDB-21		UCF101-24		Charades	
TAD	AM	TBInc	Inc	TB	GFLOPs	NPs	Val	Test	Val	Test	Val	Test
X					18.1	14,243,657	77.81	76.22	45.62	44.70	18.05	18.00
	X				16.3	13,553,560	73.96	73.75	39.26	39.01	16.82	15.40
		X			19.8	14,632,676	75.50	75.07	43.11	42.03	17.97	17.88
			X		20.5	14,356,079	75.09	74.90	42.00	41.05	17.90	17.76
				X	16.2	13,743,934	76.47	75.04	43.21	42.37	17.99	17.91
X		X			22.7	15,532,775	81.10	**81.03**	49.33	**48.85**	22.70	**22.58**
X			X		23.4	15,256,178	80.50	80.00	49.08	47.99	18.25	20.91
X				X	20.4	14,644,033	79.83	79.37	47.07	46.95	21.07	19.74
	X	X			22.2	14,842,678	77.40	77.30	46.64	45.86	18.03	18.02
	X		X		23.1	14,566,081	74.98	74.97	45.06	44.77	18.20	18.05
	X			X	17.3	13,953,936	74.65	74.55	45.04	45.00	18.00	17.97

and Inc. Meanwhile, FRH using a pair configuration of TAD and TBInc obtains the significantly highest video-mAP of 81.03%, 48.85%, and 22.58% for JHMDB-21, UCF101-24, and Charades, respectively. The best pair configuration of TAD and TBInc adopted in FRH increases action detection precisions by 4.15%–4.81% video-mAP compared to the best single configuration of FRH using only TAD on three datasets. The increased precision amounts were contributed by TBInc. Since Charades has more temporal context than JHMDB-21 and UCF101-24, the rich temporal context can be effectively leveraged by a pair configuration of TAD and TBInc in FRH, which is suitable for action detection on challenging large-scale datasets.

b: EFFECTIVENESS OF R_FSRH_D AND GASA

Structures of DNN-2 in variant simplified configurations based on FSRH, R_FSRH_D, and GASA were evaluated using video-mAP@0.5. In Table 4, the architecture of 3DCNN+RPN+R_FSRH_D+GASA achieves the best performance with increasing video-mAPs from 0.33% to 11.54%, from 1.86% to 13.95%, and from 0.66% to 9.62% for JHMDB-21, UCF101-24, and Charades, respectively. The main reason is that the architecture of 3DCNN+RPN+R_FSRH_D+GASA can provide significant spatiotemporal information as well as correlation context from an atomic subject and the other subjects for action detection in videos. The spatiotemporal semantic approach in GASA can further improve precision owing to the complementary properties between atomic subject features and subject-correlated features. The operation of GASA is an effective attention mechanism for action detection.

TABLE 4. Performance of DNN-2 at variant configurations of FSRH, R_FSRH_D and GASA on JHMDB-21, UCF101-24 and Charades.

Architectures	JHMDB-21	UCF101-24	Charades
3DCNN+RPN	73.56	41.10	16.08
3DCNN+RPN+FSRH	82.73	50.99	23.71
3DCNN+RPN+R_FSRH_D	84.77	53.19	25.04
3DCNN+RPN+R_FSRH_D +GASA	**85.10**	**55.05**	**25.70**

3) ACTION DETECTION ON AVA AND CHARADES

a: EFFECTIVENESS OF R_FSRH_D AND GASA ON AVA

Table 5 lists the simulation results of the proposed DNN-2, 3DCNN+RPN+R_FSRH_D+GASA, and the other variant architectures for action detection on AVA dataset. Compared to 3DCNN+RPN, three types of 3DCNN+RPN+FSRH, 3DCNN+RPN+R_FSRH_D, and 3DCNN+RPN+R_FSRH_D+GASA structures increased video-mAPs by 7.06%, 8.95%, and 9.70%, respectively. The inference time, second per frame, was estimated using two 1080Ti GPUs. The proposed architecture is very promising for action detection at 77ms/frame thanks to the progressive highway refinements and attention mechanisms. The proposed DNN-2 based on 3DCNN+RPN+R_FSRH_D+GASA achieves the best performance. Although action detection is more challenging

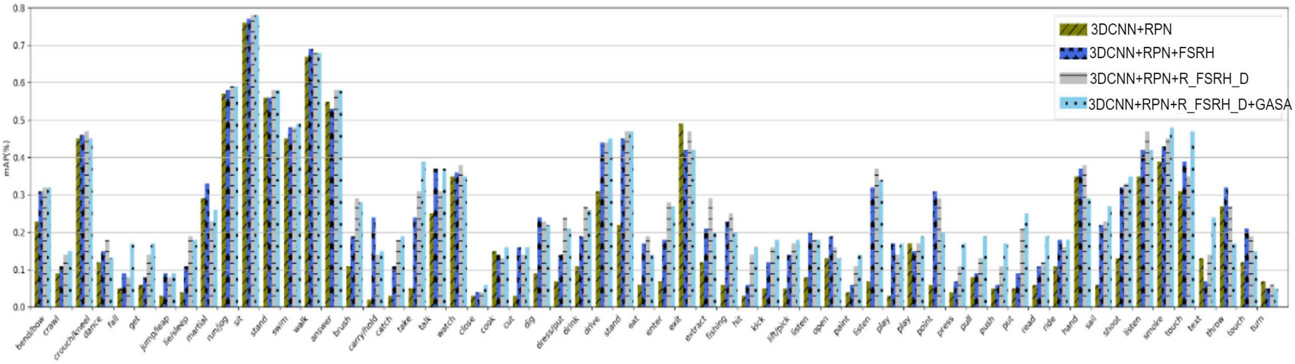


FIGURE 9. Video-mAP results associated with each category of AVA by using four architectures in Table 5.

TABLE 5. Performance of DNN-2 at variant configurations of FSRH, R_FSRH_D and GASA on AVA.

Architectures	video-mAP	Complexities	
		ms/frame	NP (M)
3DCNN+RPN	18.03	35	13,343,558
3DCNN+RPN+FSRH	25.09	59	16,943,986
3DCNN+RPN+R_FSRH_D	26.98	75	21,322,420
3DCNN+RPN+R_FSRH_D +GASA	27.73	77	22,463,742

for AVA, the proposed DNN-2 still achieves state-of-the-art result of 27.73% video-mAP at IoU@0.5.

Fig. 9 displays the video-mAP results associated with each category of AVA using 3DCNN+RPN, 3DCNN+RPN+FSRH, 3DCNN+RPN+R_FSRH_D, and 3DCNN+RPN+R_FSRH_D+GASA (DNN-2). In DNN-2, there are distinct improvements in 27 out of 60 categories in which 1% to 10% video-mAPs are increased compared to the other three counterparts. Additionally, a slight increase of 0% to 1% video-mAP occurs in 32 categories using DNN-2. The relatively large gains take place in the categories of "text on/look at a cellphone" (+10.0% video-mAP), "fall down" (+9.0% video-mAP), "take an object from a person" (+8.0% video-mAP), "push an object" (+8.0% video-mAP), "read" (+8.0% video-mAP), and "touch an object" (+8.0% video-mAP). In these categories, it is critical to characterize the subject action dynamics that are well addressed by the proposed DNN-2. Our model is only worse in one category of "turn (e.g., a screwdriver)" (-1.0% video-mAP). In summary, there are performance improvements in many categories, including multiple interactions and relations among subjects. This is because the GASA module extracts the correction features from atomic and other subjects, allowing the network to focus on the subject's behavior in the complicated surroundings.

b: DNNs WITH FAST, SLOW, OR FAST-SLOW CONFIGURATIONS IN RHs

In the proposed DNN-2, the FSRH block adopts two RHs of FRH and SRH. Accordingly, there are three topologies of FRH, SRH, and FRH+SRH (FSRH), which can be used in DNN-2. Restated, they are DNN-2 adopting R_FRH_D, R_SRH_D, and R_FSRH_D at τ =3. The experimental results at the video-mAP@0.5 metrics are summarized in Table 6. The proposed DNN-2 using R_FSRH_D noticeably outperformed the other two, with a precision improvement from 0.18% to 1.17% video-mAP on both the Charades and AVA datasets. This is because the integrated FRH and SRH can pay attention to the diverse structural features of fast and slow motions associated with subject actions.

TABLE 6. Multi-class classification performance of DNN-2 using R_FRH_D, R_SRH_D and R_FSRH_D on Charades and AVA.

Architectures	video-mAP@0.5	
	Charades	AVA
R_FRH_D	24.77	27.53
R_SRH_D	24.49	27.09
R_FSRH_D	25.66	27.71

4) STATE-OF-THE-ART COMPARISONS

We compare the proposed DNN-1 with the conventional ones listed in Tables 7 and 8 associated with action recognition tasks on JHMDB-21, UCF101-24, and TP. In Table 7, R_FSRH_D proposed in our DNN-1 is strongly complementary to the 3DCNN. State-of-the-art performance was achieved on UCF101-24 and the second best on JHMDB-21. Regarding the best outcome on JHMDB-21, the conventional DNN using the Pose Action 3D (PA3D) architecture via multi-level semantic factorization and temporal pose convolution of RGB, spatiotemporal poses, and motions, which is a complicated architecture. This shows that DNN-1 is an effective and efficient topology for action recognition. In Table 8, the proposed DNN-1 significantly outperforms the recent conventional approaches using the convolutional pose machine and the pose graph convolution network on the TP dataset [23], [24]. This indicates that the proposed DNN-1 can learn discriminative features for action recognition.

TABLE 7. Performance of DNN-1 and the other DNNs on JHMDB-21 and UCF101-24.

DNNs	Years	Pre-train	JHMDB-21	UCF101-24
T-CNN [47]	2017	Sports-1M	67.2	94.4
RPAN [9]	2017	UCF101	82.6	-
I3D [6]	2017	2017	84.1	-
PoTion+I3D [8]	2018	ImNet+Kinetics	80.9	-
RBF+RNN [48]	2018	ILSVR 2012	73.0	98.0
PA3D [7]	2019	Charades	**86.1**	**-**
CapsNet+Ske-Att [5] (reproduced)	2021	ImNet+Kinetics	85.5	98.1
Our DNN-1		ImNet+Kinetics	86.0	**98.6**

TABLE 8. Performance of DNN-1 and the other DNNs on TP.

DNNs	Years	Features	Accuracy (%)
CPM+LSTM [23]	2020	Pose+handcrafted	93.30
GCN [24]	2020	Pose Graph	97.50
Our DNN-1		Pretrained ImNet+Kinetics R_FSRH_D	**98.13**

TABLE 9. Performance of the proposed DNN-2 and conventional ones on JHMDB-21 and UCF101-24 at Video-mAP@0.5.

DNNs	Years	JHMDB-21	UCF101-24
P3D-CTN [10]	2019	80.5	-
SSD-CNNs [49]	2017	72.0	46.3
ATC [39]	2017	73.7	51.4
ACAM [50]	2020	83.9	-
Faster RCNN+I3D [20]	2018	78.6	**59.9**
YOWO[45]	2019	85.9	53.1
TPnet [51]	2018	74.1	**-**
MOC [52]	2020	77.2	54.4
Action-Tube CNN [53]	2018	-	41.1
HISAN [30]	2019	**86.4**	51.5
Dance with Flow [22]	2019	74.7	50.3
Our DNN-2		85.1	55.1

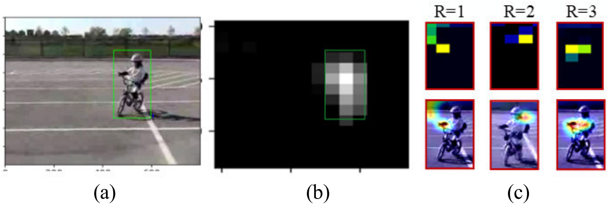


FIGURE 10. Visualization of R_FSRH_D. (a) Bounding box of a subject. (b) Feature map of a subject. (c) Pixel-wise feature maps highlighting the spatial regions at τ = 1, 3 and 5 in R_FSRH_D.

TABLE 10. Performance of the proposed DNN-2 and conventional ones on AVA and Charades.

DNNs	Years	Charades	AVA
CRF [54]	2017	22.4	-
ActionVLAD+iDT [55]	2017	21.0	-
Faster RCNN+I3D [20]	2018	15.8	-
SOA [56]	2018	16.9	-
Action Transformer[38]	2019	-	25.0
SlowFast [16]	2019	-	27.1
Long-term bank [43]	2019	-	27.2
SAGAN [17]	2020	-	23.0
ACAM [50]	2020	-	24.4
Coarse-Fine [57]	2021	25.1	-
PDAN [37]	2021	**26.5**	-
Our DNN		25.7	**27.7**

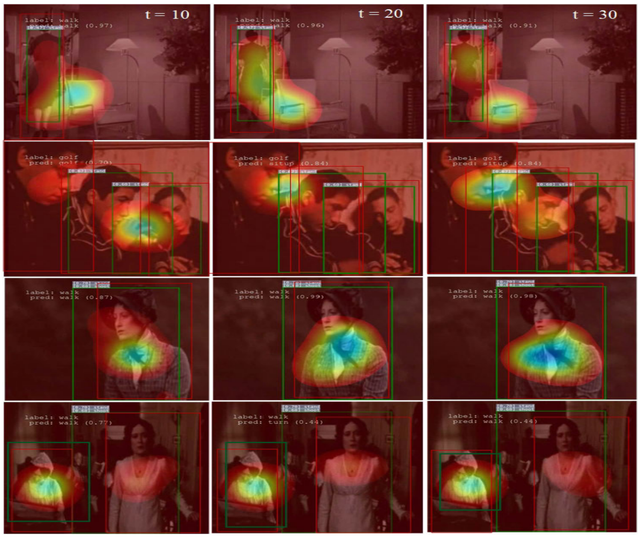


FIGURE 11. Results of DNN-2 for action detection at four activity examples of JHMDB21 where red boxes being the ground truth, and green boxes being the predicted boundary subjects associate with multiple labels and their confident scores.

The action detection results of the proposed DNN-2 and the conventional ones on JHMDB-21 and UCF101-24 are listed in Table 9. Following the standard settings, the results of action detection were obtained using the video-mAP criterion at IoU = 0.5. Overall, the proposed DNN-2 achieves comparable outcomes, in which most categories have a single action per subject. Accordingly, the mechanism of the proposed DNN-2 using GASA has not been effectively addressed. According to Tables 7 and 9, the proposed DNNs exhibit the best performance in action recognition and competitive performance in action detection, respectively. Notably, DNN-2 achieves the best compromised performance (85.1%+55.1% = 140.2%) from the two datasets when performance summation from the two datasets is considered.

Table 10 lists outcomes of the proposed DNN-2 and conventional ones on Charades and AVA. The proposed DNN-2 demonstrates state-of-the-art performance on AVA, and the second best on Charades. The proposed DNN-2 achieves 27.7% video-mAP on AVA, outperforming the second best by 0.5% video-mAP. Therefore, the proposed DNN-2 can be successfully applied to various daily-life videos for outstanding context-aware performance.

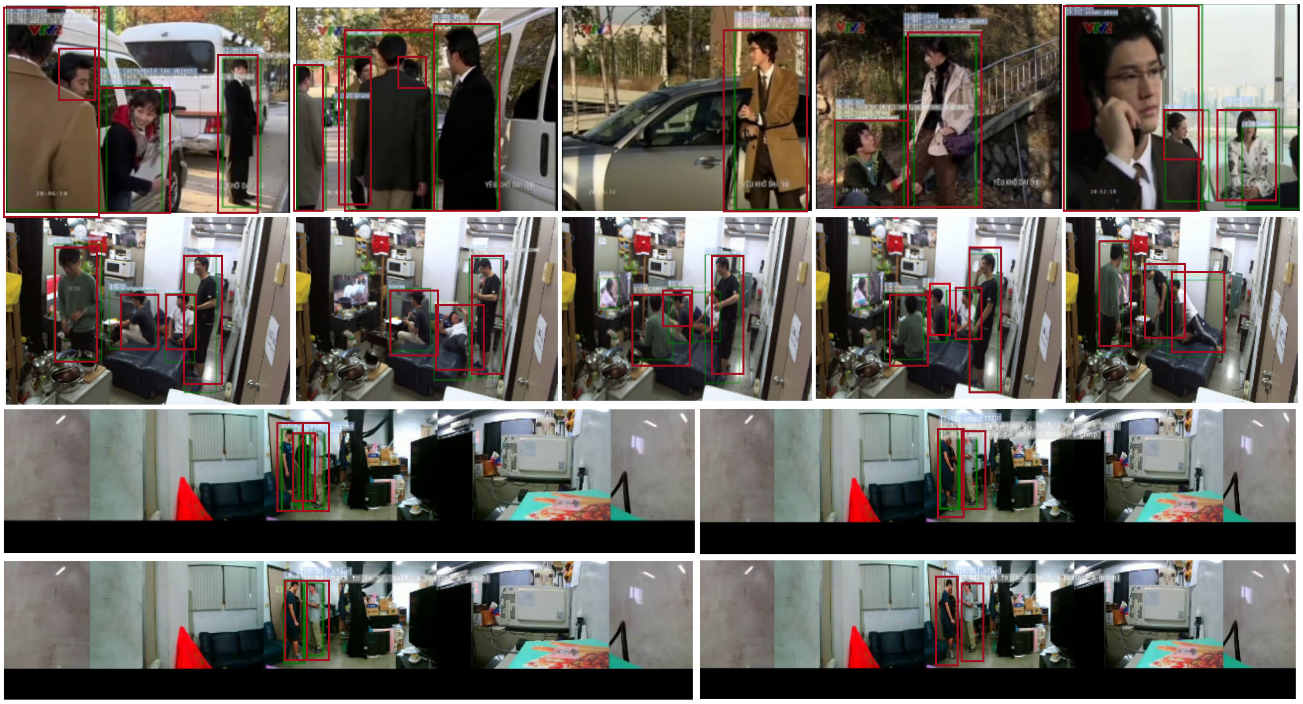


FIGURE 12. Results of DNN-2 at different video examples from AVA (1st row), Logistics-Webcam (2nd row) and panoramic camera (3rd and 4th rows), where the estimated bounding boxes marked by red and green colors being and being not the ground-truth, respectively.

5) VISUALIZATION

In this section, we qualitatively evaluate the proposed DNNs using the following visualizations. First, Fig. 10(a) shows the bounding box of a subject for action detection. The bounding box of a subject is processed to yield highlight features, as shown in Fig. 10(b). To further validate the effectiveness of the FSRH blocks in R_FSRH_D, a comparison of different numbers of τ is illustrated in Fig. 10(c). R_FSRH_D at $\tau = 3$ can effectively emphasize the critical part related to an action, revealing the effectiveness of dense contextual information aggregation for action detection. However, at $\tau = 5$, the critical part related to an action likely leaves out of the main body as compared to the case at $\tau = 3$. This phenomenon illustrates that the proposed R_FSRH_D adopts three cascaded FSRHs at $\tau = 3$.

To understand what has been learned in the process of DNN-2, we visualize the attention from R_FSRH_D and GASA. In Fig. 11, the results from DNN-2 are visualized, where four activity examples of JHMDB21 are displayed at time steps of 10, 20, and 30. A warm color reveals a high attention intensity. R_FSRH_D and GASA can highlight the meaningful parts of motions from different subjects and integrate them together as a discriminative representation for action detection. Specifically, with only one subject in a scene, the emphasized part appears in the subject body, as depicted in rows 1 and 3 of Fig. 11. When multiple subjects exist in a scene, GASA provides a relatively important clue regarding an atomic subject in rows 2 and 4 of Fig. 11. Hence, the proposed R_FSRH_D and GASA can leverage significant information from their attention mechanisms to improve action localization and prediction.

Fig. 12 shows some detection examples on the datasets from AVA, a webcam [58], and a panoramic camera [59], [60] in real indoor actions. The detection results of the proposed DNN-2 for different videos are visualized, where each subject has multiple actions. In each row, the detection outputs for a specific video clip are displayed. Orange is a false-positive localization. The bounding boxes are labeled with a red color if the detection results are ground-truth. Otherwise, the predicted bounding boxes are labeled in green. This illustrates that the proposed R_FSRH_D and GASA jointly learning in the spatial and temporal domains via the attention mechanisms is an effective way to improve the discriminative capability for action detection.

V. CONCLUSION

In this study, we successfully developed two DNNs based on R_FSRH_D and GASA to achieve outstanding performance in action recognition and detection. DNN-1 has a 3DCNN, R_FSRH_D, RL, and classification layer for action recognition. DNN-2 includes a 3DCNN, RPN, R_FSRH_D, GASA, and classification-localization layer for action detection. R_FSRH_D has three cascaded FSRHs with dense residual connections to progressively dig out significant context and long temporal information, where FSRH consists of two RHs to extract a pair of features from fast and slow actions. In particular, the proposed FSRH block with the same dimensions of input and output can be embedded in any DNN to become an inception layer for performance enhancement. GASA discovers the correlations between an atomic subject and other subjects to perceive subject interactions from an atomic subject point of view. Extensive experiments on the five video datasets were performed and examined. In action recognition, the proposed DNN-1 can achieve state-of-the-art performance of 98.6% on UCF101-24 and 98.1% on TP. In action detection, our DNN-2 achieves state-of-the-art video-mAP of 27.7% on AVA and the second-best video-mAP of 25.7% on Charades, to the best of our knowledge. Therefore, the proposed DNN-1 and DNN-2 are prominent models for the outstanding performance of action recognition and detection in various context-aware video applications.

REFERENCES

[1] A. Richard, H. Kuehne, and J. Gall, "Action sets: Weakly supervised action segmentation without ordering constraints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5987–5996.

[2] K. Soomro and M. Shah, "Unsupervised action discovery and localization in videos," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 696–705.

[3] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A hierarchical deep temporal model for group activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1971–1980.

[4] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6450–6459.

[5] 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.

[6] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6299–6308.

[7] A. Yan, Y. Wang, Z. Li, and Y. Qiao, "PA3D: Pose-action 3D machine for video recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7922–7931.

[8] V. Choutas, P. Weinzaepfel, J. Revaud, and C. Schmid, "PoTion: Pose MoTion representation for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7024–7033.

[9] W. Du, Y. Wang, and Y. Qiao, "RPAN: An end-to-end recurrent pose-attention network for action recognition in videos," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3725–3734.

[10] J. Wei, H. Wang, Y. Yi, Q. Li, and D. Huang, "P3D-CTN: Pseudo-3D convolutional tube network for spatio-temporal action detection in videos," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 300–304.

[11] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 305–321.

[12] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng, "Multi-fiber networks for video recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 352–367.

[13] J. Lin, C. Gan, and S. Han, "TSM: Temporal shift module for efficient video understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7083–7093.

[14] R. Kumar Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," 2015, *arXiv:1505.00387*.

[15] Y. Kim, Y. Jernite, D. Sontag, and A. Rush, "Character-aware neural language models," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 2741–2749.

[16] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6202–6211.

[17] M. Kim, T. Kim, and D. Kim, "Spatio-temporal slowfast self-attention network for action recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 2206–2210.

[18] R. Hou, C. Chen, and M. Shah, "An end-to-end 3D convolutional neural network for action detection and segmentation in videos," 2017, *arXiv:1712.01111*.

[19] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.

[20] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, C. Schmid, and J. Malik, "AVA: A video dataset of spatio-temporally localized atomic visual actions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6047–6056.

[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[22] J. Zhao and C. G. M. Snoek, "Dance with flow: Two-in-one stream action detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9935–9944.

[23] J. He, C. Zhang, X. He, and R. Dong, "Visual recognition of traffic police gestures with convolutional pose machine and handcrafted features," *Neurocomputing*, vol. 390, pp. 248–259, May 2020.

[24] Z. Fang, W. Zhang, Z. Guo, R. Zhi, B. Wang, and F. Flohr, "Traffic police gesture recognition by pose graph convolutional networks," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Oct. 2020, pp. 1833–1838.

[25] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 510–526.

[26] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6546–6555.

[27] X. Yang, P. Molchanov, and J. Kautz, "Making convolutional networks recurrent for visual sequence learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6469–6478.

[28] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5533–5541.

[29] H. Xu, A. Das, and K. Saenko, "R-C3D: Region convolutional 3D network for temporal activity detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5783–5792.

[30] R. R. A. Pramono, Y.-T. Chen, and W.-H. Fang, "Hierarchical self-attention network for action localization in videos," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 61–70.

[31] I. Bello, B. Zoph, Q. Le, A. Vaswani, and J. Shlens, "Attention augmented convolutional networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3286–3295.

[32] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[33] J. Li, X. Liu, W. Zhang, M. Zhang, J. Song, and N. Sebe, "Spatio-temporal attention networks for action recognition and detection," *IEEE Trans. Multimedia*, vol. 22, no. 11, pp. 2990–3001, Nov. 2020.

[34] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.

[35] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[36] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[37] R. Dai, S. Das, L. Minciullo, L. Garattoni, G. Francesca, and F. Bremond, "PDAN: Pyramid dilated attention network for action detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 2970–2979.

[38] R. Girdhar, J. Joao Carreira, C. Doersch, and A. Zisserman, "Video action transformer network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 244–253.

[39] V. Kalogeiton, P. Weinzaepfel, V. Ferrari, and C. Schmid, "Action tubelet detector for spatio-temporal action localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4405–4413.

[40] D. Li, Z. Qiu, Q. Dai, T. Yao, and T. Mei, "Recurrent tubelet proposal and recognition networks for action detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 303–318.

[41] M. Najibi, M. Rastegari, and L. S. Davis, "G-CNN: An iterative grid based object detector," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2369–2377.

[42] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.

[43] C.-Y. Wu, C. Feichtenhofer, H. Fan, K. He, P. Krahenbuhl, and R. Girshick, "Long-term feature banks for detailed video understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 284–293.

[44] Y. Li, S. Song, Y. Li, and J. Liu, "Temporal bilinear networks for video action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8674–8681.

[45] O. Köpüklü, X. Wei, and G. Rigoll, "You only watch once: A unified CNN architecture for real-time spatiotemporal action localization," 2019, *arXiv:1911.06644*.

[46] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, "A short note about kinetics-600," 2018, *arXiv:1808.01340*.

[47] R. Hou, C. Chen, and M. Shah, "Tube convolutional neural network (T-CNN) for action detection in videos," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5822–5831.

[48] Y. Shi, B. Fernando, and R. Hartley, "Action anticipation with RBF kernelized feature mapping RNN," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 301–317.

[49] G. Singh, S. Saha, M. Sapienza, P. Torr, and F. Cuzzolin, "Online real-time multiple spatiotemporal action localisation and prediction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3637–3646.

[50] O. Ulutan, S. Rallapalli, M. Srivatsa, C. Torres, and B. S. Manjunath, "Actor conditioned attention maps for video action detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 527–536.

[51] G. Singh, S. Saha, and F. Cuzzolin, "Predicting action tubes," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 106–123.

[52] Y. Li, Z. Wang, L. Wang, and G. Wu, "Actions as moving points," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 68–84.

[53] E. H. P. Alwando, Y.-T. Chen, and W.-H. Fang, "CNN-based multiple path search for action tube detection in videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 104–116, Jan. 2020.

[54] G. A. Sigurdsson, S. Divvala, A. Farhadi, and A. Gupta, "Asynchronous temporal fields for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 585–594.

[55] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell, "Action-VLAD: Learning spatio-temporal aggregation for action classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 971–980.

[56] J. Ray, H. Wang, D. Tran, Y. Wang, M. Feiszli, L. Torresani, and M. Paluri, "Scenes-objects-actions: A multi-task, multi-label video dataset," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 635–651.

[57] K. Kahatapitiya and M. S. Ryoo, "Coarse-fine networks for temporal activity detection in videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8385–8394.

[58] O. T.-C. Chen, C.-H. Tsai, H. H. Manh, and W.-C. Lai, "Activity recognition using a panoramic camera for homecare," in *Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2017, pp. 1–6.

[59] *Datasheet Brio Ultra HD Pro Business Webcam*. Accessed: Mar. 2021. [Online]. Available: https://www.logitech.com/content/dam/logitech/en_us/video-collaboration/pdf/brio-datasheet.pdf

[60] 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 *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Oct. 2020, pp. 1–5.

MANH-HUNG HA received the B.S. degree in information communication technology from the Information Communication Technology of University, Vietnam, in 2009, the M.S. degree in information communication technology from the University of Paris 13, France, in 2014, and the Ph.D. degree in electrical engineering from the National Chung Cheng University, Taiwan, in 2021. Since September 2021, he has been a Lecturer at the Faculty of Electrical Engineering, Phenikaa University, Hanoi, Vietnam. His research interests include image processing, computer vision, machine learning, speech signal processing, and knowledge engineering.

OSCAL TZYH-CHIANG CHEN (Senior Member, IEEE) received the B.S. degree in electrical engineering from the National Taiwan University, in 1987, and the M.S. and Ph.D. degrees in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 1990 and 1994, respectively. He worked with the Computer Processor Architecture Department, Computer Communication & Research Laboratories (CCL), Industrial Technology Research Institute (ITRI), working as a System Design Engineer, a Project Leader, and a Section Chief, from 1994 to 1995. He was an Associate Professor with the Department of Electrical Engineering, National Chung Cheng University (NCCU), Chiayi, Taiwan, from September 1995 to August 2003. He also worked as the Director of the Academic Development Division, Office of Research and Development, NCCU, from July 2001 to July 2004, and the Director of the Technology Transfer Center, NCCU, from July 2003 to July 2004. Since August 2003, he has been a Professor with the Department of Electrical Engineering, NCCU, where he worked as the Department Chair, from August 2018 to July 2021. He was a Visiting Scholar with the Department of Electrical and Computer Engineering, Carnegie Mellon University, USA, from December 2007 to May 2008, and the Department of Electrical and Computer Engineering, University of Wisconsin–Madison, USA, from February 2011 to July 2011. He has published more than 170 journal articles and conference papers and five book chapters, and holds 36 Taiwan patents, 22 U.S. patents, and one Chinese patent. His research interests include multimedia processing and understanding, neural networks, VLSI systems, and communication systems. He is a Life Member of the Chinese Fuzzy Systems Association. In the technical society, he was an Associate Editor of the *IEEE Circuits and Devices Magazine*, from August 2003 to December 2006, and a Founding Member of the Multimedia Systems and Applications Technical Committee of IEEE Circuits and Systems Society. He also participated in the technical program committees of many IEEE international conferences and symposiums.