Else-Net: Elastic Semantic Network for Continual Action Recognition from Skeleton Data

Tianjiao Li [1] Qiuhong Ke [2] Hossein Rahmani [3] Rui En Ho [1] Henghui Ding [4] Jun Liu [1] *
[1] Singapore University of Technology and Design [2] University of Melbourne
[3] Lancaster University [4] Nanyang Technological University
{tianjiao_li, ruien_ho}@mymail.sutd.edu.sg, jun_liu@sutd.edu.sg
h.rahmani@lancaster.ac.uk, qiuhong.ke@unimelb.edu.au, ding0093@e.ntu.edu.sg

Abstract

Most of the state-of-the-art action recognition methods focus on offline learning, where the samples of all types of actions need to be provided at once. Here, we address continual learning of action recognition, where various types of new actions are continuously learned over time. This task is quite challenging, owing to the catastrophic forgetting problem stemming from the discrepancies between the previously learned actions and current new actions to be learned. Therefore, we propose Else-Net, a novel Elastic Semantic Network with multiple learning blocks to learn diversified human actions over time. Specifically, our Else-Net is able to automatically search and update the most relevant learning blocks w.r.t. the current new action, or explore new blocks to store new knowledge, preserving the unmatched ones to retain the knowledge of previously learned actions and alleviates forgetting when learning new actions. Moreover, even though different human actions may vary to a large extent as a whole, their local body parts can still share many homogeneous features. Inspired by this, our proposed Else-Net mines the shared knowledge of the decomposed human body parts from different actions, which benefits continual learning of actions. Experiments show that the proposed approach enables effective continual action recognition and achieves promising performance on two large-scale action recognition datasets.

1. Introduction

Skeleton-based human action recognition has been attracting increasing attention in recent years because of its research significance [5, 32, 29] and relevance to a wide range of applications, e.g., human-robot interaction, self-driving vehicles, and security surveillance [34, 6]. Most of the existing works [7, 30, 38, 20, 13, 17, 26] mainly focus on using offline learning strategies to train the action recognition models, i.e., all training action sequences need to be provided at once when training the fixed-structure models. However, the recognition models operating in the real world may be exposed to continuous streams of new information, i.e., new unseen actions may continuously come in. For instance, in the real-world human-robot interaction scenario, the robot operates under open-set and can always encounter new human interactions. Given an unseen human interaction, retraining the robot on all previously observed interactions hinders the robot from conducting efficient learning and providing a timely response. In this case, the recognition models must learn from the non-stationary data distributions. However, continual learning of human action recognition under non-stationary conditions is challenging due to the catastrophic forgetting problem [23], which refers to the tendency that the recognition models forget the previously learned knowledge upon learning new unseen actions.

On the other hand, humans have an extraordinary capacity to learn continually from the external environment and their historical experience over their lifespan without the catastrophic forgetting problem, i.e., they excel at continually acquiring and accumulating new knowledge and skills. This is because human brains can learn new knowledge by searching and consolidating the most relevant memories in multiple neocortex regions or establishing new memories by activating new neocortex regions [9, 22]. In this way, human brains can turn new knowledge into long-term memories to avoid forgetting. Moreover, when learning each new knowledge, humans do not need to be retrained with all the historical information to avoid the forgetting of the old knowledge.

In this paper, we aim to investigate a brain-inspired model that can approach human intelligence for continual human action recognition, i.e., the model needs to effectively accumulate new knowledge from actions over time

[*] Corresponding Author.

while retaining the previously learned knowledge. More specifically, we propose a novel Elastic Semantic Network (Else-Net) that consists of multiple layers of elastic units. Each elastic unit comprises several learning blocks storing diversified knowledge from different human actions, with a switch block to select the most relevant block. Unlike existing offline learning approaches [7, 37] that update the parameters of the fixed-structure network during learning, our Else-Net has the capabilities in dynamically and flexibly searching and activating *only* the most relevant learning block in each Elastic Unit. It can also explore new learning blocks to store new knowledge, given the current input information. Conditioning on the selected learning blocks, our Else-Net constructs a pathway that best matches the current new human action. Since we select the learning blocks that are most relevant to the new actions for parameter updating, our network is able to learn the newly-incoming actions very effectively. Meanwhile, since the parameters of the non-selected (irrelevant) blocks are frozen, our model also preserves the knowledge of previously learned actions at the same time.

However, it can be difficult to find a matched relevant pathway for newly-incoming human actions at a holistic level, as unseen human actions may differ significantly from previously learned actions as a whole. Regardless of the overall human body, we observe that the current new action may share homogeneous features with the previously learned actions at the decomposed semantic body-part level, which benefits the searching for the relevant blocks. Inspired by this, we exploit the homogeneity by designing our Else-Net with multiple semantic branches for the decomposed multiple body parts, where each semantic branch is comprised of several layers of elastic units. Thus our network searches and activates the best-matched pathway for each semantic body part separately.

2. Related Work

Continual Learning. Continual learning aiming to continuously learn incoming new skills to approach the learning process of human intelligence in the real-world scenarios, it is an emerging yet prospective and important research direction in recent years. [8, 21, 28, 2, 1, 4, 41, 25]. Most of the existing approaches on continual learning focuses on image or object recognition. Hayes *et al.* [4] proposed to effectively replay with compressed representations, rather than original input samples. Lopez-Paz *et al.* [21] introduced gradient episodic memory to learn over temporally continuous data that alleviates forgetting, and benefits knowledge to past tasks. Pham *et al.* [25] proposed a contextual transformation network to model the task-specific features for continual learning.

Different from these works, we aim to achieve continual learning of human action recognition. To effectively handle this task, considering the semantics of human actions and structures of human bodies, we design a novel Else-Net to automatically search and update the parameters of the most relevant learning blocks in accordance with each semantic body part for the incoming new action, while freezing the parameters of the irrelevant blocks for each body part. This enables effective learning of new actions while preserving the memories of previously learned ones.

Skeleton-based Human Action Recognition. Various skeleton-based action recognition approaches [5, 29, 13, 17, 20, 18, 19, 30, 42, 15, 27, 31, 39, 14] have been proposed. Zhu *et al.* [42] proposed a deep network to recognize human activities using a regularization scheme to perceive the co-occurrences among body joints. Ke *et al.* [7] leveraged 2D convolutional neural networks (CNNs) to extract features from 3D skeleton data. Yan *et al.* [37] proposed to learn both the spatial and temporal information from skeletal data via a spatial-temporal graph convolutional network (ST-GCN).

Here we address the problem of continual action recognition in skeleton data, where the network needs to effectively and contiguously learn new types of actions over time without forgetting. A flexible Else-Net with *dynamic* pathway searching and learning is designed to handle this problem.

Dynamic Network Architecture. Our network is also relevant to dynamic network designs [33, 36, 3, 40, 24]. Wang et al. [33] designed a dynamic network, called SkipNet, that adaptively adjusts the network architecture by selecting or skipping convolutional layers based on the input data. Wu et al. [35] proposed a coarse-to-fine framework automatically adjusting and selecting the suitable network structure for feature extraction from input data, which achieves good trade-off between computational cost and accuracy. Yang et al. [40] introduced a dynamic convolutional network with dynamic width and resolution designs to deal with various computation constraints.

Differently, we design a novel Else-Net to dynamically select optimal pathways (network blocks) based on the body structure and each type of new action for better continual action recognition.

3. Elastic Semantic Network

We propose a novel Elastic Semantic Network (Else-Net) for continual learning of action recognition, where new actions need to be continuously learned over time. The proposed Else-Net is capable of effectively learning new human actions and mitigating catastrophic forgetting problems of previously learned actions. Our main idea is to construct best-matched pathways for new actions, by searching and updating only the most relevant learning blocks and exploring new learning blocks for incorporating new knowledge. Below we describe the architecture of the

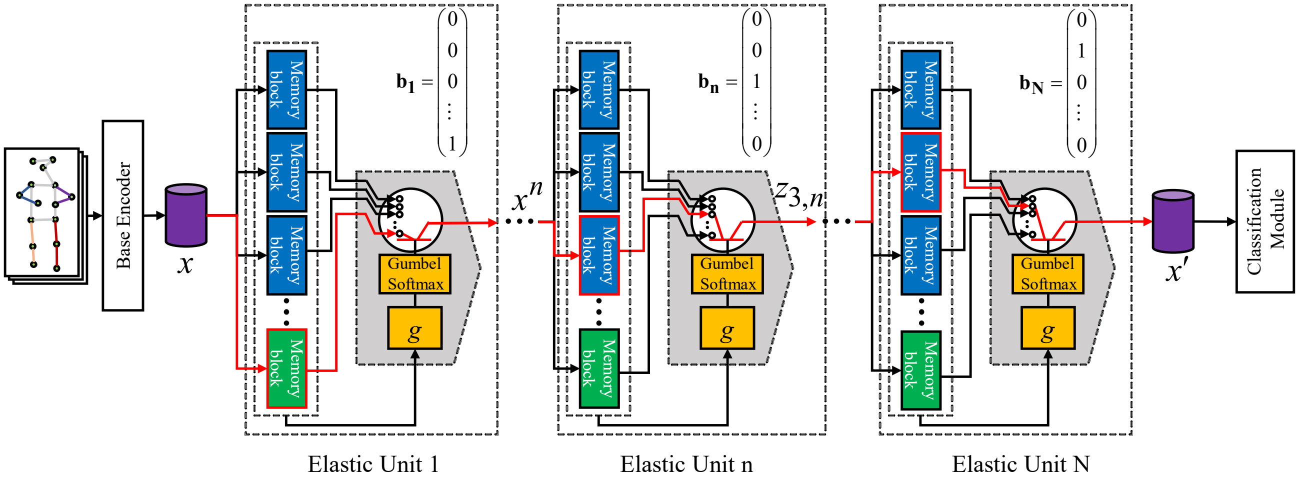

Figure 1. Illustration of the overall architecture of the proposed Else-Net. Our Else-Net is comprised of a base encoder, a stack of N elastic units and a classification module. Each skeleton sequence is first fed to a base encoder to extract the body feature x. Then, the feature passes through the elastic units, each consisting of multiple learning blocks and a switch block. The switch block in the elastic unit n selects the best-matched learning block for the input feature x^n via a Gumbel Softmax trick to produce a one-hot matching vector $\mathbf{b_n} = [b_1, \cdots, b_{B_n+1}]^T$. By conditioning on the selected learning block in each Elastic Unit, our Else-Net constructs the most relevant pathway (indicated by the red arrows), i.e., blocks on this pathway are selected to learn the current new action and to produce the latent features x' for predicting the label of the current new action. Note that only the parameters of the selected learning blocks are updated, while the non-selected blocks are frozen. Therefore, our Else-Net preserves the knowledge of the previously learned human actions as well.

proposed Else-Net in detail.

3.1. Block Search and Pathway Construction

As shown in Fig. 1, the proposed Else-Net contains a stack of N elastic units, where each elastic unit n ($n \in \{1, ..., N\}$) is comprised of several learning blocks and a switch block. We denote the learning blocks within the n^{th} elastic unit as: $\{f_{\theta_{i,n}}(\cdot)\}_{i=1}^{B_n}$, where $\theta_{i,n}$ are the parameters of the i^{th} learning block, and B_n is the number of learning blocks in this elastic unit. These learning blocks contain diversified prior knowledge that is attained from previously learned actions. The switch block, which consists of a gating module g followed by the Gumbel Softmax, is responsible for selecting the most relevant learning block within the n^{th} elastic unit according to the input feature (x^n) of this elastic unit.

Considering that there could be new knowledge in the current unseen action to be learned, an additional new learning block $f_{\theta_{B_n+1,n}}(\cdot)$ (indicated as the green block in Fig. 1) is temporarily appended to the existing blocks within the n^{th} elastic unit, as a candidate learning block for storing new knowledge. More specifically, the input feature x^n of the n^{th} elastic unit is fed to all the learning blocks (including the additional learning block) within this elastic unit, i.e., $\{f_{\theta_{i,n}}(\cdot)\}_{i=1}^{B_n+1}$, to produce the corresponding latent features: $\{z_{i,n}\}_{i=1}^{B_n+1}$. As shown in Fig. 1, the encoded latent features are then passed through the gating module g followed by the Gumbel Softmax to produce a one-hot vector $\mathbf{b_n} = [b_1, \cdots, b_{B_n+1}]^T$, where the highest score (i.e., score 1) corresponds to the best-matched block. It is worth mentioning that the additional new learning block will be permanently added to the elastic unit if it obtains the highest matching score. Otherwise, it will be removed. This process can be formulated as follows:

$$z_{i,n} = f_{\theta_{i,n}}(x^n), \quad i \in \{1, ..., B_n + 1\}, \tag{1}$$

$$\mathbf{b_n} = \text{Gumbel\_Softmax}(g(z_{1,n}); g(z_{2,n}); ...; g(z_{B_n+1,n})). \tag{2}$$

Note that if the input feature (x^n) of the n^{th} elastic unit best matches the i^{th} learning block, the output value $g(z_{i,n})$ is expected to be higher than other blocks. This is because the convolutional operations in the learning blocks are able to capture the shared homogeneous features between the current input feature x^n and the corresponding learning block $f_{\theta_{i,n}}(\cdot)$, i.e., the more correlated x^n and $f_{\theta_{i,n}}(\cdot)$ is, the higher the output value $g(f_{\theta_{i,n}}(x^n))$. Then, with the Gumbel Softmax trick, the learning block with the highest output value attains a matching score equal to 1. This block is "activated" as the best-matched block in accordance with the input feature x^n. Hence, the encoded feature ($z_{i,n}$) from this best-matched block is used as the output of the n^{th} elastic unit. By selecting the most relevant learning block within each Elastic Unit to encode the current input feature, our Else-Net can exploit the homogeneity between the current input feature and the previously learned knowledge stored in the selected learning blocks to achieve effective learning of the current new action.

Moreover, utilizing the Gumbel Softmax to generate the one-hot matching vector also ensures that only the parameters of the most relevant learning blocks are updated, while the parameters of the non-selected learning blocks, storing irrelevant knowledge learned from other actions, are frozen. This selective updating scheme enables our Else-Net to try to preserve the knowledge of the previously learned actions as much as possible and mitigate the catastrophic forgetting problem. The selective updating can be formulated as follows:

$$\theta_{i,n} \leftarrow \theta_{i,n} - \alpha \nabla_{\theta_{i,n}} [-b_{i,n} \cdot y_k \log \hat{y}_k], \tag{3}$$

where $i \in \{1, ..., B_n + 1\}$, $\theta_{i,n}$ are the parameters of the i^{th} learning block in the n^{th} elastic unit, y_k and $\hat{y}_k$ denote the ground-truth label and predicted label of the k^{th} newly incoming action sample, $b_{i,n}$ denotes the matching score (i.e., 1 for the best-matched learning block, and 0 for others) of the i^{th} learning block within the n^{th} elastic unit for the k^{th} sample, and α is the learning rate.

Note our Else-Net contains N elastic units (N levels). In such a multi-level structure, different human actions can share common learning blocks at different levels, instead of exhaustively adding an additional block to every elastic unit. As mentioned above, each elastic unit is capable of selecting the most relevant learning block in it w.r.t. its input feature. Thus, by connecting all the selected relevant learning blocks of the N elastic units, a promising semantic pathway that best matches the current input features x, is constructed.

More specifically, the red arrows in Fig. 1 illustrate the best-matched semantic pathway to exploit the homogeneity between x and the previously learned knowledge, i.e., x flows through the most relevant learning blocks (shown in red boxes) sequentially, to mine the homogeneous features and produce the latent semantic features x'. Hence, our Else-Net is able to learn the newly-incoming actions very effectively by utilizing and strengthening the most relevant blocks and prior knowledge, while well retaining the knowledge stored in the non-selected irrelevant blocks.

3.2. Pathway Construction for Body Part Branches

Although the proposed Else-Net described in Sec. 3.1 dynamically constructs and updates the best-matched pathway, given the full skeleton, it is still quite challenging to select an optimal pathway that best matches the newly incoming action. This is because the new actions as a whole may differ significantly from the previously learned actions. However, these actions can still share some homogeneous knowledge at the decomposed body-part level. Inspired by this, the proposed Else-Net is further designed to be capable of mining and strengthening the shared knowledge at the decomposed body-part level, i.e., achieving effective learning of new actions via leveraging previously learned knowledge for each body part.

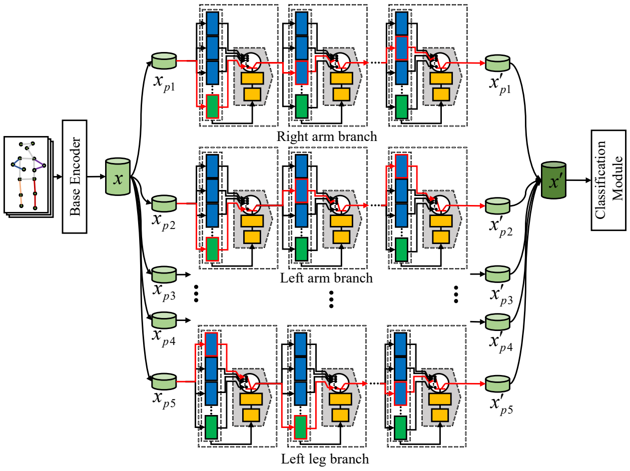

Figure 2. Illustration of the Else-Net with five branches for decomposed five local body parts. To better exploit homogeneous features, we divide the full body into five local parts. Specifically, the current new input skeletons are fed to the base encoder to extract body feature x. Given the full body feature x, we divide it to five part features, and then feed each part feature to the corresponding branch for feature learning. Then encoded latent part features are concatenated into an integral global feature (x') for action classification.

Concretely, we divide the input body feature x into five semantic body part features $\{x_{p^j}\}_{j=1}^{5}$, as shown in Fig. 2. The decomposed semantic body part features are then fed to the corresponding body part branches, i.e., left arm, left leg, body trunk, right leg, and right arm. These five branches (without parameter sharing) have the same architecture, and each branch contains N elastic units for processing the features of each body part. As introduced in Sec. 3.1, our Else-Net dynamically searches and constructs a semantic pathway that best matches the current input features. Therefore, by inheritance, each body part branch is capable of searching for the best-matched learning block in each Elastic Unit, for the input features of each corresponding body part.

With this, each body part branch constructs an optimal semantic pathway w.r.t. the current input semantic part features x_{p^j} ($j \in \{1, ..., 5\}$), and produces informative latent features x'_{p^j} ($j \in \{1, ..., 5\}$). We utilize five optimal pathways to effectively learn the input semantic features ($\{x_{p^j}\}_{j=1}^{5}$) separately. Finally, we can obtain powerful integral body features (x') by concatenating the learned features $\{x'_{p^j}\}_{j=1}^{5}$ and achieve effective learning of the current new action while mitigating the catastrophic forgetting problem.

3.3. Training and Testing

Training. Following previous continual learning settings [4, 21], we learn new human actions once and one by one,

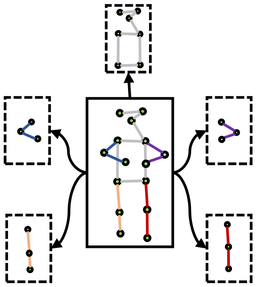

Figure 3. Illustration of decomposed five local body parts (i.e., trunk, left/right hand, and left/right leg).

i.e., each new human action is considered as a new task in continual learning settings. To train our Else-Net, we update the parameters by minimizing the classification loss (categorical cross-entropy). The optimization of our Else-Net comprises of two phases: outer optimization and inner optimization. The outer optimization fixes parameters of all learning blocks while updating parameters of switch blocks in the elastic units. The inner optimization fixes parameters of the switch blocks while updating parameters of the selected learning blocks. Specifically, when learning each new action, we first apply the outer optimization and then apply the inner optimization. The optimization procedure can be formulated as follows:

$$\text{Outer:}\quad \boldsymbol{\theta}_{\mathbf{g}} \leftarrow \boldsymbol{\theta}_{\mathbf{g}} - \alpha \nabla_{\boldsymbol{\theta}_{\mathbf{g}}}[-y_k \log \hat{y}_k] \tag{4}$$

$$\text{Inner:}\quad \boldsymbol{\theta}_{\mathbf{m}} \leftarrow \boldsymbol{\theta}_{\mathbf{m}} - \alpha \nabla_{\boldsymbol{\theta}_{\mathbf{m}}}[-\mathbf{b} \cdot y_k \log \hat{y}_k] \tag{5}$$

where $\boldsymbol{\theta}_{\mathbf{m}}$ and $\boldsymbol{\theta}_{\mathbf{g}}$ denote the parameters of learning blocks and gating modules respectively, $\mathbf{b}$ denotes the matching scores of learning blocks, and y_k and $\hat{y}_k$ denote the ground-truth label and predicted label for the k^{th} newly incoming action sample.

Testing. During inference phase, input human actions are first fed to the base encoder to extract different semantic body part features. Then, given the decomposed body part features, each semantic pathway automatically searches for the most relevant learning blocks to mine homogeneity between the current new features and the previously learned ones. Finally, the encoded body part features are concatenated and passed through the classification module to attain global features for action classification.

3.4. Implementation Details

Network Architecture. Considering the powerful capability in representing human skeleton data by disentangling multi-scale aggregation scheme to capture powerful skeleton features, we leverage MS-G3D [20] as our base encoder. Each learning block $f_{\theta_{i,n}}(\cdot)$ consists of a convolutional layer with 1×1 kernel, and ReLU activation. Note that the learning block retains the shape of input features. The gating module $g(\cdot)$ in the switch blocks is comprised of a linear layer and $tanh$ activation function to normalize the output value. Our Else-Net contains three layers of elastic units ($N = 3$). Each elastic unit is initialised with three learning blocks, and the number of learning blocks can dynamically increase during continual learning.

Body Feature Decomposition. As mentioned above, MS-G3D [20] is used as our base encoder to extract human skeleton features $x \in \mathcal{R}^{C\times V}$, where C denotes feature size for each skeleton joint and V is the total number of skeleton joints. Note that the V skeleton joints spatially correspond to the V-dimensional extracted body features. Therefore, as shown in Fig. 3, we can divide the whole body feature x into five decomposed semantic body part features $\{x_{p^j}\}_{j=1}^{5}$ along the spatial dimension.

Classification Module. The classification module is comprised of two fully-connected layers. The latent semantic body part features $\{x'_{p^i}\}_{i=1}^{5}$ encoded by the five semantic pathways are first concatenated, then fed to these fully connected layers to predict the action label.

Episodic Memory. Following the continual learning setting [4, 25, 21], we use a small episodic memory storing a small portion of observed data (10%) for replay. When new human actions are fed to the network, they are used to populate the episodic memory simultaneously. For each training iteration, two random action samples in the episodic memory are used for replay when learning a new human action.

4. Experiments

We evaluate the proposed Else-Net on two large-scale 3D skeleton-based human action recognition datasets: NTU RGB+D dataset [29] and PKU-MMD dataset [16]. The experiments are conducted on a Nvidia RTX 3090 graphics card. The initial learning rate α is set to 10^{-3}. Following the continual learning setting [4, 21], new human actions are divided into multiple tasks by their classes. At each time, only one task is used to train the recognition model for 5 iterations and this task will not be presented to the model again, i.e., the multiple tasks are learned sequentially and each task is generally observed only once, except for those stored in the small episodic memory. Following the setting in [4] where some categories are learned continuously while the others are used for pre-training to provide the model with prior knowledge, for all continual learning settings we continuously learn 10 new unseen categories, while the other categories are used for pre-training. For offline learning settings, we follow MS-G3D [20] and train 50 epochs over all action samples.

4.1. Datasets

NTU RGB+D [29] is a large-scale action recognition dataset, widely used for 3D skeleton action recognition.

The NTU RGB+D dataset contains 60 action classes and 56,880 videos. Note that this dataset contains 40 human subjects and diversified human actions, which thus are very likely to lead to forgetting of old actions, when the model continually learns new human actions. The NTU RGB+D dataset provides two standard evaluation protocols, namely cross-view (CV) and cross-subject (CS). In the CS protocol, sequences performed by 20 subjects are used for training, and remaining sequences are used for testing. In the CV protocol, sequences captured from 2 viewpoints are used for training, and remaining sequences are used for testing.

PKU-MMD dataset [16] is a large 3D skeleton dataset, providing 51 action classes and 1,076 untrimmed videos, containing 21,545 labelled action instances performed by 66 distinct subjects. The evaluation protocols of PKU-MMD are similar to those of NTU RGB+D, i.e., a cross-view (CV) protocol where 2 viewpoints are used for training and the remaining viewpoint is used for testing, and a cross-subject (CS) protocol where action videos of 57 subjects are used for training and the remaining videos are used for testing.

4.2. Evaluation Criteria

To evaluate the abilities of the proposed Else-Net in both effectively learning the new actions and mitigating catastrophic forgetting, we follow the metrics introduced in [25] for continual learning performance evaluation. The metrics include Average Accuracy (ACC), Forgetting Measure (FM) and Learning Accuracy (LA). Suppose that our model aims to learn a total of T action classes sequentially, and $a_{t,q}$ represents the recognition accuracy on action class q after the model is trained on the action class t.

Average Accuracy (ACC) is defined as the average recognition accuracy of all observed actions after training the model on the last action (i.e., action class T):

$$ACC = \frac{1}{T}\sum_{q=1}^{T} a_{T,q} \tag{6}$$

Forgetting Measure (FM) evaluates how much knowledge has been forgotten after the model has been trained continually up till action class T. The lower the FM, the more unlikely the model forgets the previously learned actions. The forgetting measure is formulated as:

$$FM = \frac{1}{T-1}\sum_{q=1}^{T-1} \max_{t\in\{1,2,\dots,T-1\}} \{a_{t,q} - a_{T,q}\} \tag{7}$$

Learning Accuracy (LA) evaluates the recognition performance of the model on an action class immediately after training on this action, which reflects the model's ability in learning current new actions.

$$LA = \frac{1}{T}\sum_{q=1}^{T} a_{q,q} \tag{8}$$

4.3. Experimental Results on PKU-MMD

We compare the proposed Else-Net with the state-of-the-art continual learning approaches [21, 4] for the task of continual action recognition on the PKU-MMD dataset. To ensure fair comparison across the continual learning experiments, we fix the learning order of the incoming actions, i.e., the models are trained continually on the same sequence of action classes. We also use the proposed Else-Net to conduct offline action recognition (i.e., all actions can be accessed at once), and compare with the state-of-the-art approaches of skeleton-based action recognition for offline learning. The results are shown in Table 1.

Results on Continual Learning. Our Else-Net achieves the best performance across all metrics on both cross-subject and cross-view evaluation protocols, compared to existing continual learning approaches [4, 21]. The significant improvement demonstrates that the proposed Else-Net can exploit the homogeneous features between current new actions and previously learned human actions, effectively learning new knowledge and preserving past knowledge. Following the continual learning settings where some categories are learned sequentially while others are used for pre-training, we sequentially learn 10 categories and other categories are used for pre-training. During testing phase, all the categories are used to evaluate our model.

It is worth mentioning that, when compared to existing continual learning approaches, such as GEM [21] and Remind [4], our Else-Net achieves significant improvement on FM (lower the better). This shows that by selecting and updating the most relevant learning blocks, our Else-Net preserves past knowledge and thus avoid forgetting previous human actions. Also, as shown in Table 1, our Else-Net attains higher LA, which shows its ability to effectively learn new incoming tasks in a continual learning approach. This means that the optimal semantic pathways constructed using the block searching strategy is able to mine homogeneous features between current input actions and previously learned human actions, which empowers the proposed model to learn new actions effectively.

Besides, we conduct an experiment on training our backbone encoder, i.e., MS-G3D [20] in the continual learning manner. The significant performance improvement of our network over MS-G3D further demonstrates that the proposed block search and body part pathway construction scheme has the capabilities in effectively learning new human actions and mitigating catastrophic forgetting.

Results on Offline Learning. To further evaluate the capacities of our model, we evaluate our method under the offline learning setting, where all action categories can be accessed at once. The results of offline learning are shown in Table 1. The proposed Else-Net achieves state-of-the-art performance over other approaches. This indicates that although our Else-Net is specifically designated for contin-

Table 1. Performance comparison (%) on PKU-MMD. Our model trained under the continual learning setting outperforms other continual learning methods, and even achieves competitive results compared to models trained under the offline learning setting. Besides, under the offline learning setting used by previous skeleton-based action recognition methods, we also obtain competitive performance.

Setting	Methods	CS			CV		
		ACC	FM	LA	ACC	FM	LA
Continual Learning	GEM [21]	65.9	13.5	72.8	61.3	12.7	74.3
	Remind [4]	71.2	7.5	85.1	75.3	8.7	81.3
	MS-G3D [20]	65.3	17.0	77.2	68.0	23.7	72.7
	Else-Net	**84.6**	**4.0**	**86.8**	**87.0**	**7.2**	**90.8**
Offline Learning	Li *et al.* [10]	90.4	-	-	93.7	-	-
	HCN [11]	92.6	-	-	94.2	-	-
	RF-Action [12]	92.9	-	-	94.4	-	-
	MS-G3D [20]	93.1	-	-	94.9	-	-
	Else-Net	**95.3**	-	-	**97.2**	-	-

Table 2. Performance comparison (%) on NTU RGB+D

Setting	Methods	CS			CV		
		ACC	FM	LA	ACC	FM	LA
Continual Learning	GEM [21]	55.3	15.1	72.1	54.5	11.5	64.7
	Remind [4]	56.0	9.5	66.5	59.8	9.4	68.9
	MS-G3D [20]	46.3	25.4	56.4	54.5	23.1	58.5
	Else-Net	**84.4**	**5.1**	**87.6**	**87.9**	**8.0**	**89.3**
Offline Learning	ST-GCN [37]	81.5	-	-	88.3	-	-
	2s-AGCN [4]	88.5	-	-	95.1	-	-
	MS-G3D [20]	91.5	-	-	96.2	-	-
	Else-Net	**91.6**	-	-	**96.4**	-	-

ual learning, the block searching and body part pathway construction strategy is also beneficial for offline learning-based action recognition.

4.4. Experimental Results on NTU RGB+D

We conduct extensive experiments in continual learning and offline learning settings on the very challenging NTU RGB+D dataset to evaluate the efficacy of the proposed network. For fair comparisons, the order of the action tasks is fixed across all continual learning experiments. During training phase, we continuously learn 10 categories and the others are used for pre-training to provide prior knowledge to the model. And during inference, all categories are used for evaluation.

Results on Continual Learning. To evaluate the efficacy of our model, we compare the proposed Else-Net with existing continual learning approaches. As shown in Table 2, the proposed Else-Net outperforms existing approaches on ACC and LA by large margins, demonstrating that selecting the most relevant learning blocks for the current input human action enables the model to exploit homogeneity between the newly-input and previously learned knowledge. In addition, our Else-Net achieves lower FM compared to other approaches. This demonstrates that by updating the parameters of the selected relevant blocks while freezing the non-selected ones, our Else-Net preserves past knowledge stored in the non-selected learning blocks to mitigate the forgetting problem.

Results on Offline Learning. As shown in Table 2, we also compare our Else-Net with offline learning approaches. The proposed Else-Net achieves competitive performance compared to existing offline learning methods. This further demonstrates the capability of our Else-Net, that is specifically designed for continual action recognition, even on action recognition under the offline learning setting.

4.5. Ablation Study

Below, we conduct extensive ablation experiments on NTU RGB+D dataset (cross-subject protocol) to evaluate the efficacy of our proposed Else-Net from different perspectives.

Impacts of Number of Elastic Units. To evaluate the impact of the number of Elastic Units (EU), we conduct experiments on NTU RGB+D dataset (cross-subject protocol) by stacking different numbers of EUs in each semantic pathway. As shown in Tab. 3, with the growth of number of Elastic Units, the ACC and LA increase while the FM decreases. This could be explained as that increasing number of Elastic Units empowers our model with more representative capacities in mining the homogeneous knowledge between the current new human actions and the previously-learned actions and achieving better continual learning performance. We also observe that when the number of EUs ≥ 3, the increase in performance becomes trivial for all metrics, including ACC, FM and LA. In Tab. 3, compared to 3 Elastic Units, when Else-Net is comprised of 4 Elastic Units, the ACC and LA increase only 0.2% and 0.1% respectively, while the FM decreases only 0.1%. This shows that three Elastic Units could be qualified for learning sufficient representative features to effectively learn new human actions and avoid forgetting past learned actions.

Table 3. Performance comparison (%) of different numbers of Elastic Units (EUs) in each branch.

Num. of EUs	ACC	FM	LA
1	80.2	7.7	84.0
2	83.2	6.3	86.6
3	84.4	5.1	87.6
4	84.6	5.0	87.7

Impact of Body Part Branches. To evaluate the efficacy of employing decomposed body part branches to exploit homogeneous features, we compare our method illustrated in Fig. 2 with the method illustrated in Fig. 1, where whole body features are used as input features instead of decomposing into semantic part features. Specifically, we use a single branch for full-body feature learning, i.e., **"Else-Net w/o Part Branches"**, and report results in Table 5. For a fair comparison, we initialize the model with single body branch and the model with five body-part branches with the same number of learning blocks.

The experimental results in Table 5 demonstrate that the recognition performance drops and the forgetting measure FM increases, when the body-part branches are replaced by a single full-body branch. This is because there may be significant discrepancies between the current new action

Table 4. Growth rate (GR) of the number of learning blocks, when the model is trained until the final task.

Methods	GR	ACC	FM	LA
Else-Net w/o Part Branches	1070%	83.0	7.4	86.0
Else-Net	**68%**	**84.4**	**5.1**	**87.6**

and previously learned actions. Thus, the single-branch network for full human body is limited in exploiting homogeneity between actions for effective learning of new actions by taking advantage of prior knowledge of other actions. On the contrary, different body part branches are able to boost the learning of current new human action by separately exploiting the semantic homogeneous features from different local body parts, where the shared homogeneity with previously learned human actions are more likely to be observed, which can be exploited to learn the new action effectively.

Moreover, compared to the single-branch network for processing the full body, our body-part branches have a lower growth rate (GR) as shown in Table 4. Specifically, the growth rate is calculated as $\frac{\Delta}{\mathbb{N}}$, where Δ denotes the increased number of learning blocks (i.e., the number of new learning blocks that are added) during continual learning from the first task until the final task, and $\mathbb{N}$ denotes the number of initial learning blocks before the model starts to learn the first task.

As shown in Table 4, the growth rate of single branch model is more than ten times larger than our Else-Net, which means that instead of searching for the relevant blocks with homogeneous knowledge learned from previous actions, the single branch model tends to more exhaustively explore brand new learning blocks when new tasks coming in. However, our Else-Net is able to make better use of the previously-learned learning blocks while moderately exploring new blocks, and thus it is able to achieve a good recognition performance, by dynamically searching and updating the most relevant learning blocks to effectively learn newly-incoming human actions and preserve the past knowledge from the previous human actions.

Impact of Block Searching. We conduct experiments to evaluate the impact of block searching and pathway construction. We leverage the same number of pre-defined learning blocks for different Else-Net variations as the number of initial learning blocks. Instead of selecting the most relevant block in each elastic unit, all latent features produced by all learning blocks in the elastic units are concatenated and sent into a fully-connected layer to fuse the information from all learning blocks. Then, the fused features are fed into next learning block sequentially. The experimental results of this setting (**"Else-Net w/o Block Searching"**) are shown in Table 5.

Table 5. Ablation Study (%) on NTU RGB+D (CS)

Methods	ACC	FM	LA
Else-Net w/o Block Searching	76.4	17.5	83.5
Else-Net w/o Selective Updating	77.3	13.4	84.2
Else-Net w/o Part Branches	83.0	7.4	86.0
Else-Net	**84.4**	**5.1**	**87.6**

We analyse that the performance discrepancy of this setting (without using block searching) compared to our Else-Net, may come from two aspects: 1) By concatenating all the encoded latent features together, the diversified knowledge from different learning blocks are mixed-up. Thus, the unwanted irrelevant noise may harm the recognition performance; 2) Since all learning blocks are used and updated, the past knowledge stored in learning blocks may be overwritten, leading to forgetting problems. Conversely, our Else-Net dynamically searches and updates the most relevant blocks in accordance with the current input actions to achieve effective learning, while preserving past knowledge by freezing irrelevant blocks.

Impact of Selective Updating. To evaluate the efficacy of selective updating of the selected learning blocks, we replace the Gumbel Softmax function with a Softmax function in the switch block, and call this variant **"Else-Net w/o Selective Updating"**. In this case, the matching scores of all blocks are non-zero, i.e., all the learning blocks in the elastic units are updated w.r.t. current input features. As shown in Table 5, the performance drops when we adopt the variant of "Else-Net w/o Selective Updating". We analyse that updating the parameters of all the learning blocks disturbs past knowledge from previously learned human actions. This also introduces irrelevant noise when the model learns new human actions, leading to a performance drop.

5. Conclusion

In this paper, we propose a brain-inspired Elastic Semantic Network, namely Else-Net, for continual human action recognition. The proposed Else-Net is able to dynamically search for the most relevant learning blocks with regard to the input human actions and exploit the homogeneous features between the current new actions and the previously learned human actions to achieve effective learning of current new human action. In addition, our Else-Net is able to selectively update the parameters of the most relevant learning blocks, while freezing non-selected learning blocks to preserve previously learned knowledge, in order to mitigate catastrophic forgetting problems. An optimal semantic pathway is further constructed on the selected relevant learning blocks to mine the homogeneity over each decomposed local body part. With such a block search and body-part pathway construction process, the proposed Else-Net shows great efficacy in learning new human actions and preserving the old knowledge from previously learned human actions.

Acknowledgement This work is supported by AISG-100E-2020-065, SUTD Projects PIE-SGP-AI2020-02, SRG-ISTD-2020-153 and the TAILOR project funded by EU Horizon 2020 programme under GA No 952215.

References

[1] Rahaf Aljundi, Lucas Caccia, Eugene Belilovsky, Massimo Caccia, Min Lin, Laurent Charlin, and Tinne Tuytelaars. Online continual learning with maximally interfered retrieval. *arXiv preprint arXiv:1908.04742*, 2019. 2

[2] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*, 2018. 2

[3] Cristobal Eyzaguirre and Alvaro Soto. Differentiable adaptive computation time for visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12817–12825, 2020. 2

[4] Tyler L Hayes, Kushal Kafle, Robik Shrestha, Manoj Acharya, and Christopher Kanan. Remind your neural network to prevent catastrophic forgetting. In *European Conference on Computer Vision*, pages 466–483. Springer, 2020. 2, 4, 5, 6, 7

[5] Jian-Fang Hu, Wei-Shi Zheng, Jianhuang Lai, and Jianguo Zhang. Jointly learning heterogeneous features for rgb-d activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5344–5352, 2015. 1, 2

[6] Jian-Fang Hu, Wei-Shi Zheng, Lianyang Ma, Gang Wang, and Jianhuang Lai. Real-time rgb-d activity prediction by soft regression. In *European Conference on Computer Vision*, pages 280–296. Springer, 2016. 1

[7] Qiuhong Ke, Mohammed Bennamoun, Hossein Rahmani, Senjian An, Ferdous Sohel, and Farid Boussaid. Learning latent global network for skeleton-based action prediction. *IEEE Transactions on Image Processing*, 29:959–970, 2019. 1, 2

[8] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 2

[9] Dharshan Kumaran, Demis Hassabis, and James L McClelland. What learning systems do intelligent agents need? complementary learning systems theory updated. *Trends in cognitive sciences*, 20(7):512–534, 2016. 1

[10] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Skeleton-based action recognition with convolutional neural networks. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 597–600. IEEE, 2017. 7

[11] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. *arXiv preprint arXiv:1804.06055*, 2018. 7

[12] Tianhong Li, Lijie Fan, Mingmin Zhao, Yingcheng Liu, and Dina Katabi. Making the invisible visible: Action recognition through walls and occlusions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 872–881, 2019. 7

[13] Tianjiao Li, Jun Liu, Wei Zhang, and Lingyu Duan. Hard-net: Hardness-aware discrimination network for 3d early activity prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 2

[14] Tianjiao Li, Jun Liu, Wei Zhang, Yun Ni, Wenqian Wang, and Zhiheng Li. Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16266–16275, 2021. 2

[15] Wenbo Li, Longyin Wen, Ming-Ching Chang, Ser Nam Lim, and Siwei Lyu. Adaptive rnn tree for large-scale human action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1444–1452, 2017. 2

[16] Chunhui Liu, Yueyu Hu, Yanghao Li, Sijie Song, and Jiaying Liu. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. *arXiv preprint arXiv:1703.07475*, 2017. 5, 6

[17] Jun Liu, Amir Shahroudy, Mauricio Lisboa Perez, Gang Wang, Ling-Yu Duan, and Alex Kot Chichung. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 1, 2

[18] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European conference on computer vision*, pages 816–833. Springer, 2016. 2

[19] Jun Liu, Gang Wang, Ping Hu, Ling-Yu Duan, and Alex C Kot. Global context-aware attention lstm networks for 3d action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1647–1656, 2017. 2

[20] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 143–152, 2020. 1, 2, 5, 6, 7

[21] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *arXiv preprint arXiv:1706.08840*, 2017. 2, 4, 5, 6, 7

[22] James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419, 1995. 1

[23] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989. 1

[24] Yue Meng, Chung-Ching Lin, Rameswar Panda, Prasanna Sattigeri, Leonid Karlinsky, Aude Oliva, Kate Saenko, and Rogerio Feris. Ar-net: Adaptive frame resolution for efficient action recognition. In *European Conference on Computer Vision*, pages 86–104, 2020. 2

[25] Quang Pham, Chenghao Liu, Doyen Sahoo, and Steven HOI. Contextual transformation networks for online continual learning. In *International Conference on Learning Representations*, 2021. 2, 5, 6

[26] Hossein Rahmani, Arif Mahmood, Du Huynh, and Ajmal Mian. Histogram of oriented principal components for

cross-view action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 38(12):2430–2443, 2016. 1

[27] Hossein Rahmani, Ajmal Mian, and Mubarak Shah. Learning a deep model for human action recognition from novel viewpoints. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):667–681, 2017. 2

[28] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017. 2

[29] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1010–1019, 2016. 1, 2, 5

[30] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12026–12035, 2019. 1, 2

[31] Yansong Tang, Yi Tian, Jiwen Lu, Peiyang Li, and Jie Zhou. Deep progressive reinforcement learning for skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5323–5332, 2018. 2

[32] Xionghui Wang, Jian-Fang Hu, Jian-Huang Lai, Jianguo Zhang, and Wei-Shi Zheng. Progressive teacher-student learning for early action prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3556–3565, 2019. 1

[33] Xin Wang, Fisher Yu, Zi-Yi Dou, Trevor Darrell, and Joseph E Gonzalez. Skipnet: Learning dynamic routing in convolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 409–424, 2018. 2

[34] Junwu Weng, Xudong Jiang, Wei-Long Zheng, and Junsong Yuan. Early action recognition with category exclusion using policy-based reinforcement learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(12):4626–4638, 2020. 1

[35] Zuxuan Wu, Caiming Xiong, Yu-Gang Jiang, and Larry S Davis. Liteeval: A coarse-to-fine framework for resource efficient video recognition. In *Advances in Neural Information Processing Systems*, 2019. 2

[36] Zuxuan Wu, Caiming Xiong, Chih-Yao Ma, Richard Socher, and Larry S Davis. Adaframe: Adaptive frame selection for fast video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1278–1287, 2019. 2

[37] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 2, 7

[38] Siyuan Yang, Jun Liu, Shijian Lu, Meng Hwa Er, and Alex C Kot. Collaborative learning of gesture recognition and 3d hand pose estimation with multi-order feature analysis. In *European Conference on Computer Vision*, pages 769–786. Springer, 2020. 1

[39] Siyuan Yang, Jun Liu, Shijian Lu, Meng Hwa Er, and Alex C Kot. Skeleton cloud colorization for unsupervised 3d action representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 2

[40] Taojiannan Yang, Sijie Zhu, Chen Chen, Shen Yan, Mi Zhang, and Andrew Willis. Mutualnet: Adaptive convnet via mutual learning from network width and resolution. In *European Conference on Computer Vision*, pages 299–315, 2020. 2

[41] Huaxiu Yao, Yingbo Zhou, Mehrdad Mahdavi, Zhenhui Li, Richard Socher, and Caiming Xiong. Online structured meta-learning. *arXiv preprint arXiv:2010.11545*, 2020. 2

[42] Wentao Zhu, Cuiling Lan, Junliang Xing, Wenjun Zeng, Yanghao Li, Li Shen, and Xiaohui Xie. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016. 2