View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Lancaster E-Prints

Global Regularizer and Temporal-aware Cross-entropy for Skeleton-based Early Action Recognition

Qiuhong Ke[1], Jun Liu[2], Mohammed Bennamoun[1], Hossein Rahmani[3],
Senjian An[4], Ferdous Sohel[5], and Farid Boussaid[1]

[1] The University of Western Australia, Crawley, Australia
[2] Nanyang Technological University, Singapore
[3] Lancaster University, Lancashire, England
[4] Curtin University, Bentley, Australia
[5] Murdoch University, Murdoch, Australia
qiuhong.ke@research.uwa.edu.au, jliu029@ntu.edu.sg,
mohammed.bennamoun@uwa.edu.au, h.rahmani@lancaster.ac.uk,
s.an@curtin.edu.au, f.sohel@murdoch.edu.au, farid.boussaid@uwa.edu.au

Abstract. In this paper, we propose a new approach to recognize the class label of an action before this action is fully performed based on skeleton sequences. Compared to action recognition which uses fully observed action sequences, early action recognition with partial sequences is much more challenging mainly due to: (1) the global information of a long-term action is not available in the partial sequence, and (2) the partial sequences at different observation ratios of an action contain a number of sub-actions with diverse motion information. To address the first challenge, we introduce a global regularizer to learn a hidden feature space, where the statistical properties of the partial sequences are similar to those of the full sequences. We introduce a temporal-aware cross-entropy to address the second challenge and achieve better prediction performance. We evaluate the proposed method on three challenging skeleton datasets. Experimental results show the superiority of the proposed method for skeleton-based early action recognition.

Keywords: Early action recognition · Global regularizer · Temporal-aware cross-entropy · 3D skeleton sequences.

1 Introduction

Early action recognition, which aims to infer the action class before an action is fully performed, is very important for a wide range of real-world applications, such as prevention of dangerous events in video surveillance, autonomous driving and health care systems [1]. It can also be used to enhance human-robot interaction, allowing robots to quickly respond to humans and improve user experience [2]. Given that human actions are performed in the 3D space, 3D skeleton sequences captured by depth cameras provide comprehensive and useful information for action analysis [3]. 3D skeleton data is also robust to clustered

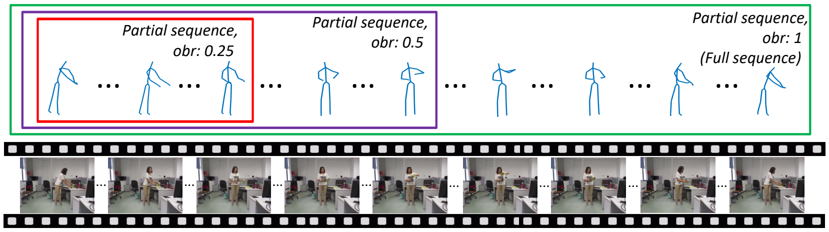


Fig. 1. A human action generally contains multiple temporal stages and sub-actions. Early action recognition aims to recognize the action class before the action sequence is fully observed, i.e., the observation ratio ('obr') is smaller than 1.

backgrounds and illumination variations. Besides, such data is easy to obtain due to the prevalence and affordability of 3D sensors. Moreover, the dimension of the skeleton data is quite small (75 in each frame in the skeleton sequence captured by Kinect V2). This makes skeleton data very attractive for real-time applications (such as early action recognition). In this paper, we propose a new early action recognition approach based on 3D skeleton data.

A series of advanced skeleton-based action recognition methods have been proposed in recent years and have shown promising performance [4–10]. Compared to action recognition, early action recognition is much more challenging. In action recognition, an action is recognized after the action is fully observed. In contrast, in early action recognition, the goal is to infer the action class when the action is still in progress. In other words, the observation used for early action recognition is often a partial sequence, which only contains a partial execution of the action. As a result, the long-term global information of the action, which is important for class inference, is unavailable. As shown in Figure 1, a human action may last for a long period of time and contain multiple sub-actions, with a variety of human motions during the overall duration of the action. As a result, in order to accurately infer the action class, the long-term global information of the full action progress needs to be captured and modeled [11–13].

In this paper, we propose a global regularizer to learn a latent feature space, in which the partial sequences and the full sequences are similar. More specifically, during training, both the partial and the full sequences are fed to the network as two inputs, which are then aggregated in a hidden feature layer by using the global regularizer. The global regularizer measures the distance of the statistical properties between the partial and full sequences in the latent space. We propose to minimize the discrepancy of their statistical properties (i.e., mean and variance) rather than directly minimizing their pair-wise feature distance. This is because the estimation of the statistical properties can help reduce the effects of outliers such as actions without motions occurring at an early temporal stage. Once the network is trained, the learned network (which shares information of the full sequences) is used to process a given testing sequence to infer the action class.

Another main challenge in early action recognition is that the partial sequences of an action class often contain diverse motion information. As shown in Figure 1, the partial sequence with a small observation ratio only contains some early sub-actions, while the partial sequence with a large observation ratio contains more information of the action. Intuitively, the partial sequence with a small observation ratio should be given a smaller weight during the training of the network as less action information is provided. To address this issue, we introduce a temporal-aware cross-entropy as the classification loss of the network. The temporal-aware cross-entropy is calculated based on the observation ratios of the partial sequences. More specifically, less penalty is given to the classification errors made by the partial sequences with smaller observation ratios. This prevents the network from over-fitting the partial sequences with small observation ratios and improves the prediction performance.

To sum up, in this paper, we address skeleton-based early action recognition. We explore two main challenges in this task, and propose a new method to handle these challenges. We summarize the main contributions of this paper as follows: (1) We propose a new network which contains a global regularizer to exploit the global information. (2) We introduce a temporal-aware cross-entropy as the classification loss for early action recognition at different observation ratios. (3) We report extensive experimental results demonstrating the superior performance of the proposed method.

2 Related Works

Skeleton-based Action Recognition Human action recognition and early recognition are important tasks in the area of computer vision due to their wide ranges of applications [14–22]. Due to the prevalence of affordable and highly-accurate sensing cameras, many efforts have been made on 3D skeleton-based action recognition [4–10, 23–26]. Traditional methods include spatial feature learning from each frame of the sequence (*e.g.*, pairwise relative positions, rotations and translations) and temporal modelling of the features of all frames using Fourier Temporal Pyramid (FTP) [4, 5]. Recently, LSTM has been developed to learn the spatial or spatial-temporal information from skeleton sequences [6, 7, 9]. Deep CNN has also been leveraged to process skeleton sequences as images, which hierarchically learns both the spatial and temporal information [10]. Different from the aforementioned works on action recognition by using fully-observed skeleton sequences, in this paper, we propose a new framework for the challenging skeleton-based early action recognition task.

Early Action Recognition Compared to action recognition (which uses full sequences for action inference), early action recognition, aims to infer the action class at an early temporal stage before the action is fully observed. For RGB-based early action recognition, most of the existing works try to exploit the temporal dynamics of the features of partial sequences [2, 27–30]. In [30],

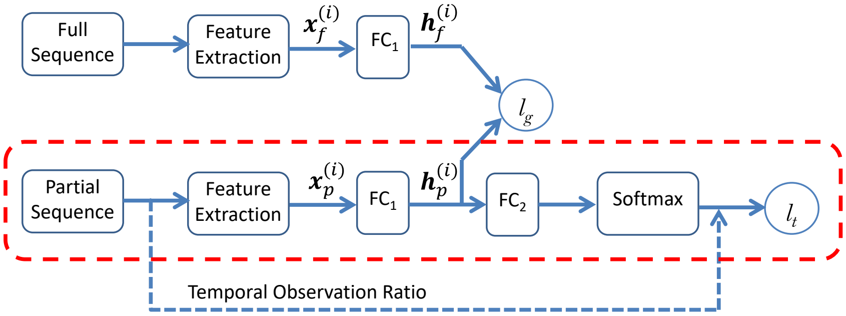


Fig. 2. The overall architecture of the proposed method. FC denotes fully connected layer. $\mathbf{x}_f^{(i)}$ and $\mathbf{x}_p^{(i)}$ denote the features of the full and partial sequences in the i^{th} training pair, respectively. $\mathbf{h}_f^{(i)}$ and $\mathbf{h}_p^{(i)}$ denote the hidden representations of the full and partial sequences, respectively. Given a testing sequence, the sub-network shown in the red box is leveraged to process the sequence for early action recognition. The full sequence and the temporal observation ratio value are not required in the testing phase.

each partial sequence is represented with spatial, temporal and structural features for early interaction recognition. Some works [31, 32] focus on learning a classification model using a new loss for better early action recognition. In [31], a weighted cross-entropy is introduced for early action recognition. The new loss aims to prevent the network from over-fitting the partial sequences at an early stage. In [32], a weighted false positive classification loss is introduced and added to the standard cross-entropy as a new classification loss for action anticipation. Recently, Farha *et al.* [33] introduced two different architectures based on CNN and RNN for future action anticipation. For RGB-D early action recognition, Hu *et al.* [34] introduced a soft regression-based framework for early action recognition using RGB-D videos. Liu *et al.* [35] introduced a Scale Selection Network (SSNet) to effectively select different scales for online 3D early action recognition. In this work, we focus on early action recognition from skeleton sequences that contain a single activity.

3 Proposed Method

The overall architecture of the proposed framework is shown in Figure 2. It contains a global regularizer, which aims to exploit the global information from partial sequences for better early action recognition. The classification loss of the network is a temporal-aware cross-entropy, which prevents the network from over-fitting the partial sequences with small observation ratios and further improves the performance. In this section, we describe the architecture in details.

3.1 Global Regularization

A human action generally contains multiple sub-actions and a high variation of human motions. The long-term global information needs to be leveraged in order to recognize the action more accurately. In early action recognition, the partial sequence does not provide the global information. This makes early action recognition very challenging. In this section, we introduce a new method to exploit the global information for early action recognition from the partial sequence. The main idea is to learn a hidden feature space, in which the partial sequences are similar to the full sequences. This similarity is enforced by minimizing a global regularizer between the full and the partial sequences in the feature space. The global regularizer is calculated based on the statistical properties of the features of the sequences. The discrepancy of the statistical properties has been shown to lead to an incompatibility between two distributions of data in domain adaptation tasks [36–39].

As shown in Figure 2, the network has two inputs. During training, the training data is fed to the network in the form of pairs. Each pair consists of the feature of the full sequence of an action sample and the feature of a partial sequence of the same sample. Here, we denote the set of training pairs as $\{\mathbf{x}_f^{(i)}, \mathbf{x}_p^{(i)}\}_{i=1}^{n}$, where n denotes the number of training pairs in the batch. $\mathbf{x}_f^{(i)} \in \mathbb{R}^{d_1}$ and $\mathbf{x}_p^{(i)} \in \mathbb{R}^{d_1}$ are the features of the full and the partial sequences in the i^{th} training pair, respectively. The features of the partial sequence and the corresponding full sequence in each pair are fed to the hidden layer $\mathtt{FC_1}$ of the network to generate an embedding feature. We use $W \in \mathbb{R}^{d_2 \times d_1}$ and $\mathbf{b} \in \mathbb{R}^{d_2 \times 1}$ to denote the weight matrix and bias vector of the hidden layer, where d_2 is the dimension of the hidden space. The hidden representations of the full and partial sequences in the $i^{(th)}$ pair are calculated as:

$$\begin{aligned} \mathbf{h}_f^{(i)} &= g(W\mathbf{x}_f^{(i)} + \mathbf{b}) \\ \mathbf{h}_p^{(i)} &= g(W\mathbf{x}_p^{(i)} + \mathbf{b}) \end{aligned} \tag{1}$$

where g is the rectified linear unit (ReLU) activation function [40].

The global regularizer l_g between the partial and the full sequences is computed as:

$$\ell_g = \left\| \hat{\mathbf{h}}^p - \hat{\mathbf{h}}^f \right\|_2 . \tag{2}$$

where $\hat{\mathbf{h}}^p$ and $\hat{\mathbf{h}}^f$ are the global representations of the partial and full sequences in the hidden space, respectively. $\hat{\mathbf{h}}^p$ and $\hat{\mathbf{h}}^f$ are calculated as:

$$\hat{\mathbf{h}}_p = \begin{bmatrix} \mathbf{m}_p \\ \mathbf{v}_p \end{bmatrix} \tag{3}$$

$$\hat{\mathbf{h}}_f = \begin{bmatrix} \mathbf{m}_f \\ \mathbf{v}_f \end{bmatrix} \tag{4}$$

where

$$\begin{aligned}\mathbf{m}_p &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\mathbf{h}_p^{(i)}\\ \mathbf{v}_p &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\left(\mathbf{h}_p^{(i)}-\mathbf{m}_p\right)^2\\ \mathbf{m}_f &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\mathbf{h}_f^{(i)}\\ \mathbf{v}_f &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\left(\mathbf{h}_f^{(i)}-\mathbf{m}_f\right)^2\end{aligned} \tag{5}$$

3.2 Temporal-aware Cross-entropy

The network simultaneously minimizes the global regularizer and the classification loss of the partial sequences, thus the sequences do not lose their discriminative power, which is used for recognizing actions. The standard cross-entropy between the ground-truth action label and the output predictions is generally used for action recognition. Specifically, let $\mathbf{z}^{(i)} = [z_1^{(i)}, \cdots, z_m^{(i)}]^T \in \mathbb{R}^m$ denote the feature vector of the i^{th} sample that is fed to the Softmax layer to generate the probability scores, where m denotes the number of action classes. The predicted probability of the k^{th} class is defined as:

$$p_k^{(i)} = \frac{\exp(z_k^{(i)})}{\sum\limits_{j=1}^{m}\exp(z_j^{(i)})} \tag{6}$$

The cross-entropy between the predicted probability and the ground-truth label is formulated as:

$$\ell_c^{(i)} = -\sum_{k=1}^{m} y_k^{(i)} \log\left(p_k^{(i)}\right) \tag{7}$$

where $y_k^{(i)}$ is the ground-truth action label corresponding to the k^{th} class of the i^{th} sample, i.e., assume the action class of the i^{th} sample is $m^\star$, then $y_k^{(i)} = 1$ if $k = m^\star$, and $y_k^{(i)} = 0$ otherwise.

In contrast to action recognition, where the testing data comes with full sequences which cover the full progress of the action, the testing data in early action recognition are partial sequences with various observation ratios. Intuitively, a partial sequence with a small observation ratio should be assigned a smaller weight during the training of the network, as the partial sequence with a small observation ratio often contains less action information and more noisy information (that is irrelevant to the full action) compared to a partial sequence with a large observation ratio. Therefore, we introduce a temporal-aware cross-entropy loss function, which uses different weights that are related to the observation ratios of the partial sequences. Specifically, this loss function gives smaller penalties to the classification mistakes of the partial sequences with smaller observation ratios. This prevents the network from over-fitting. The temporal-aware cross-entropy is formulated as:

$$\ell_t^{(i)} = f(r^{(i)})\ell_c^{(i)} \tag{8}$$

where $r^{(i)} = \frac{t}{T}$ denotes the observation ratio of a partial sequence with t frames. T denotes the total number of frames of the full sequence that the partial sequence belongs to. $\ell_c^{(i)}$ is the standard cross-entropy formulated as Equation 7 and $f(r^{(i)})$ is an increasing function. We empirically evaluated two increasing functions, including the linear form:

$$f(r^{(i)}) = r^{(i)} \tag{9}$$

and the exponential form:

$$f(r^{(i)}) = e^{r^{(i)}-1} \tag{10}$$

Both methods give smaller weights to the losses of the partial sequences with smaller observation ratios and yield better performances compared to the method without the temporal-aware cross-entropy. In the experiments, the exponential form is used to report our final prediction accuracies.

3.3 Network Training and Action Inference

The loss of the network includes the temporal-aware cross-entropy and the global regularizer:

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n} \ell_t^{(i)} + \lambda \ell_g \tag{11}$$

where n denotes the number of training pairs of the partial and full sequences, and λ is the weight to balance the two losses.

During training, the full sequences, which contain the fully executions of actions, are provided by each dataset. For each full sequence, we generate multiple partial sequences. We denote a full skeleton sequence as $s(1:J, 1:T)$, where J denotes the number of human skeletal joints in each frame, and T denotes the number of frames. We segment $s(1:J, 1:T)$ into k partial sequences which all start from the first frame. The i^{th} partial sequence can be denoted as $s(1:J, 1:\left[\frac{T \cdot i}{k}\right])$. In this case, the observation ratio of this partial sequence is $\frac{i}{k}$. Each partial sequence contains the accumulative information that starts from the first frame of the action. Consequently, the k^{th} partial sequence is actually the full sequence containing T frames.

To feed the training data to the network, we first adapt the method in [10] to extract features from the partial and full skeleton sequences. The main idea of [10] is to transform skeleton sequences into image-based representations to extract CNN features, which hierarchically encode both the spatial and temporal information. In [10], each skeleton sequence is represented as four images to extract four features. In this paper, we aim to show the benefits of the proposed framework, rather than the feature representations of the skeleton sequences. Therefore, during training and testing, we represent each sequence as only one image and extract one feature vector for computation efficiency, rather than extracting multi-features which are computational expensive as [10]. During testing, the sub-network (shown in the red box in Figure 2) is used to process a given partial sequence. The full sequence and the temporal observation ratio of the partial sequence are not required to predict the action.

4 Experiments

The proposed method was evaluated on three benchmark skeleton datasets, i.e., NTU Dataset [7], CMU Dataset [41] and SYSU 3D Human-Object Interaction (3DHOI) Dataset [42]. In this section, we report the experimental results and present detailed analysis.

4.1 Datasets

NTU Dataset [7] contains more than 56000 sequences and 4 million frames. This is currently the largest skeleton-based action dataset. There are 60 action classes, including both one-person actions (e.g., eating, drinking) and two-person interactions (e.g., kicking, handshaking). These actions are performed by 40 distinct subjects and are captured by three cameras. The cameras are placed at different locations and view points, which results in 80 view points in total. The 3D coordinates of 25 joints are provided for each skeleton. This dataset is very challenging due to the large view-point, intra-class and sequence-length variations.
CMU Dataset [41] contains 2235 sequences and about 1 million frames. This dataset has been categorized into 45 action classes (*e.g.*, walking, jumping) [8]. Each action is performed by one person. The 3D coordinates of 31 joints are provided in each skeleton. There are high sequence-number variations and intra-class diversity, which make this dataset very challenging.
SYSU 3DHOI Dataset [42] contains 480 sequences performed by 40 subjects. There are 12 action classes, including drinking, pouring, calling with a cell phone, playing with a cell phone, wearing a backpack, packing a backpack, sitting in a chair, moving a chair, taking out a wallet, taking something out from the wallet, mopping and sweeping. Some actions are very similar at the early temporal stage since the subjects operate the same object, or the actions have the same starting sub-action, such as standing still. This makes this dataset very challenging for early action recognition.

4.2 Experimental Settings

The following methods are implemented based on the same feature for comparison:
Action Recognition Network (ARN). This network is trained using only full sequences. The network has one input and one output with a Softmax layer to generate class scores. In testing, the partial sequence is directly fed to the

network for early action recognition. This baseline is used to show the importance of using partial sequences to train the early action recognition network.

Early Action Recognition Network (EARN). This baseline has the same architecture as the ARN baseline, except that the training data consists of all the available partial and full sequences. Compared to our proposed method, EARN does not include the global regularizer for early action recognition. Besides, it uses the standard cross-entropy loss for the classifier.

Early Action Recognition Network + Global Regularization (EARN + GR). In this baseline, a network with the same architecture as the proposed network is devised, except that the classification loss is the standard cross-entropy loss function. In other words, this method does not take the temporal observation ratio into account to compute the temporal cross-entropy. The loss of the network is the combination of the global regularizer and the standard cross-entropy.

Early Action Recognition Network + Global Regularization + Temporal-aware Cross-entropy (EARN + GR + TCE). This method is the proposed method, which includes the global regularizer and the temporal-aware cross-entropy for early action recognition.

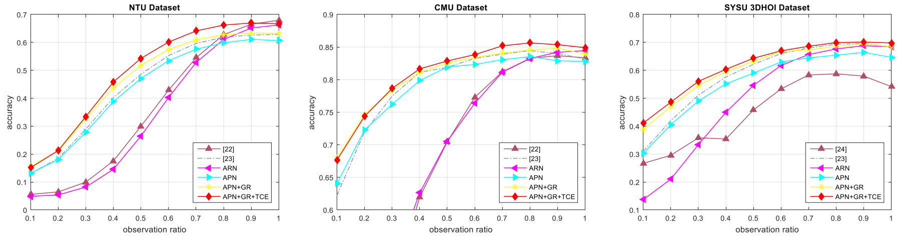
Other Early Action Recognition Methods. We also compare with other state-of-the-art early action recognition methods, including the methods in [32], [31] and [34].

In our implementation, the parameter λ in Equation 11 is set to 1 for all datasets. The number of units of the first and second fully-connected layer is set to 512 and the number of action classes, respectively. The network is trained using the stochastic gradient descent algorithm. The learning rate is set to 0.001 with a decay rate of 10^{-6}. The momentum is set to 0.9. For all datasets, we split 20% of the training data for validation to select the hyper-parameters. Each action sequence is divided into 10 partial sequences. All the partial sequences start from the first frame of the full sequence. The observation ratios of the partial sequences, in this case, are changed from 0.1 to 1, with a step of 0.1.

4.3 Results on the NTU Dataset

Table 1. Performance comparison of early action recognition on the NTU Dataset. We adapt the method in [10] to transform each skeleton into an image and extract one feature vector. The same feature is used in all methods for fair comparisons. Refer to Figure 3(a) for more results.

Observation	Methods					
Ratios	[32]	[31]	ARN	EARN	EARN+GR	EARN+GR+TCE
0.3	28.94%	9.95%	8.21%	27.87%	32.66%	**33.46%**
0.5	48.62%	29.98%	26.50%	47.13%	51.79%	**54.28%**
0.8	61.86%	62.80%	61.06%	59.88%	62.76%	**66.21%**
Average	45.15%	36.41%	34.51%	43.77%	47.27%	**49.41%**



(a) (b) (c)

Fig. 3. Performance comparison of early action recognition on (a) NTU Dataset, (b) CMU Dataset and (c) SYSU 3DHOI Dataset. On each observation ratio r, the partial sequence that starts from the first frame to the $(rT)^{th}$ frame is used for testing. T denotes the number of frames in the full sequence. (Best viewed in color)

Table 2. Performance comparison of the proposed method without and with temporal-aware cross-entropy (linear and exponential) on the NTU Dataset.

Observation Ratios	EARN+GR	EARN+GR +TCE (linear)	EARN+GR +TCE (exp)
0.3	32.66%	32.63%	**33.46%**
0.5	51.79%	53.94%	**54.28%**
0.8	62.76%	65.66%	**66.21%**
Average	47.27%	48.88%	**49.41%**

For the evaluation on this dataset, we follow the cross-subject testing protocol, i.e., the sequences of 20 subjects are used for training and the sequences of the other 20 subjects are used for testing. The comparisons of the proposed method to other methods are shown in Figure 3(a) and Table 1.

EARN outperforms ARN when the observation ratio is smaller than 0.8. More specifically, when the observation ratio is 0.5, i.e., only the former 50% of the number of frames of each sequence is used for testing, the prediction accuracy of EARN is 47.13%, which is 20.63% better than ARN (26.5%). ARN outperforms EARN when the observation ratio is close to 1. The recognition accuracy (observation ratio is 1) of ARN is 66.25%. Compared to EARN (60.64%), the improvement of ARN is 5.61%. In contrast to EARN, ARN does not use partial sequences to train the network. It clearly shows that early action recognition is different from action recognition and the partial sequences are indispensable for network training in order to predict actions at the early temporal stage.

EARN + GR is seen to significantly outperform EARN on all observation ratios. Particularly, when the observation ratio is 0.5, the prediction accuracy of the EARN + GR is 51.79%, which is 4.66% better than that of the EARN (47.13%). The improvement of the EARN + GR compared to the EARN averaged across all observation ratios is 3.5% (from 43.77% to 47.27%). EARN does not take into account the fact that the global information is not available in the partial sequences. In contrast, the proposed EARN + GR explicitly investigates the global information with a global regularizer. The global regularizer encourages the network to learn a hidden feature space, in which the partial and full sequences are similar. During testing, because the partial sequences are mapped to the hidden space which shares the global information of the full sequences, better performances are achieved. It clearly shows the benefits of the global regularizer for early action recognition.

EARN + GR + TCE further improves the performance of EARN + GR, especially at the late temporal stage. When the observation ratio is 0.5, the performance of the EARN + GR + TCE is 54.28%. Compared to the EARN + GR (51.79%), the performance of EARN + GR + TCE is 2.49% better. The improvement of EARN + GR + TCE compared to EARN + GR averaged across all observation ratios is 2.14% (from 47.27% to 49.41%). EARN + GR uses standard cross-entropy for the classification of the partial sequences. In contrast, EARN + GR + TCE takes the temporal observation ratios of the sequences into account and calculates the temporal cross-entropy as the classification loss of the partial sequences. The temporal cross-entropy prevents the network from over-fitting the sequences with small observation ratios and yields better performances.

We also compared the proposed method with state-of-the-art early action recognition methods [31, 32] in Figure 3 and Table 1. We used the same feature in [31, 32] for fair comparisons. When the observation ratio is 0.5, the performance of [32] is 48.62%. EARN + GR + TCE outperforms [32] by 5.66% (from 48.62% to 54.28%). The average improvement of the proposed EARN + GR + TCE compared to [32] is 4.26%. In [32], a new loss combining the standard cross-entropy and a linearly increasing false positive loss, is introduced to prevent ambiguities in early action recognition. It does not explicitly exploit the global information for early action recognition. The EARN + GR + TCE also significantly outperforms the method in [31]. More specifically, the performance of EARN + GR + TCE (54.28%) is 24.30% better than that of the method in [31] (29.98%) on observation ratio 0.5. In [31], a weighted cross-entropy is introduced for early action recognition. The weight of the cross-entropy of each partial sequence is an exponential function in terms of the difference of the number of the frames between the partial and the full sequences. It can be seen that the performance of this method is similar to ARN (which is trained only on the full sequences). This is due to the fact that the numbers of frames of most full sequences are large, which yields zero weights for most partial sequences. In other words, the network does not take the partial sequences for training. This results in a degraded performance of early action recognition at the early temporal stage due to the under-fitting problem. The method in [31] performs better only when the number of frames of the testing sequence is similar to that of the full sequence (i.e., observation ratio is close to 1) as this method mainly focuses on full sequences for training.

Table 2 compares the performances of the proposed method with linear (Equation 9) and exponential (Equation 10) cross-entropy on the NTU Dataset.

We refer to these two methods as EARN + GR + TCE (linear) and EARN + GR + TCE (exp). Both methods outperform EARN + GR, which uses standard cross-entropy instead of the temporal-aware cross-entropy for classification. The improvements of EARN + GR + TCE (linear) and EARN + GR + TCE (exp) compared to EARN + GR averaged across all observation ratios are 1.61% and 2.14%, respectively. EARN + GR + TCE (exp) outperforms EARN + GR + TCE (linear) on all observation ratios. EARN + GR + TCE (linear) uses the observation ratio of each partial sequence as the weight to calculate cross-entropy. This method suppresses the weight of the partial sequence with a small observation ratio to a very low value. In other words, the network does not train the partial sequences with small observation ratios, which results in low prediction accuracies during testing due to the under-fitting problem. With EARN + GR + TCE (exp), the weights of the partial sequences range from 0.37 ($\exp^{0-1}$) to 1 ($\exp^{1-1}$). In this case, the network trains the partial sequences of all observation ratios.

4.4 Results on the CMU Dataset

For this dataset, we follow the 4-fold cross-validation protocol proposed in [8] for evaluation. In particular, we use the four different splits of the data provided by [8] for training and testing. The average accuracies of the four folds are reported. The comparison results of the proposed method with the baselines and the state-of-the-art early action recognition methods [31, 32] are shown in Figure 3(b) and Table 3.

Similar to the NTU Dataset, the performance of ARN is better than EARN only when the observation ratio is close to 1. The proposed EARN + GR achieves an average accuracy of 80.55%, which outperforms EARN (78.9%) by 1.65%. EARN + GR + TCE further improves the performance of EARN + GR to 81.01%. Compared to the method in [32] and [31], the improvements of the proposed EARN + GR + TCE are by 1.67% and 18.01%, respectively. The improvements of the proposed method are smaller on this dataset compared to that on the NTU Dataset. This is due to the fact that most actions of this dataset (*e.g.*, running and jumping) are periodical and are repeated many times across the full sequences. In this case, the motions at different temporal stages are similar and the information of the partial sequences is similar to the global information of the full sequences. Therefore, the actions can be accurately recognized even without the global information, which limits the benefits of global regularizer and temporal-aware cross-entropy.

4.5 Results on the SYSU 3DHOI Dataset

For evaluation, we followed the cross-subject setting used by [34], in which the data samples performed by half of the subjects were used for training, and the data samples from the other half were used for testing. The accuracies were averaged over 30 different combinations of training and testing subjects provided by [42]. The results are shown in Figure 3(c) and Table 4. In this paper we only

Table 3. Performance comparison of early action recognition on the CMU Dataset. The same feature is used in all methods for fair comparisons. Refer to Figure 3(b) for more results.

Observation	Methods					
Ratios	[32]	[31]	ARN	EARN	EARN+GR	EARN+GR+TCE
0.3	77.45%	47.50%	50.91%	76.22%	78.34%	**78.65%**
0.5	81.87%	70.45%	70.53%	81.95%	82.56%	**82.87%**
0.8	84.45%	83.26%	83.25%	83.53%	84.56%	**85.63%**
Average	79.34%	63.00%	65.21%	78.90%	80.55%	**81.01%**

Table 4. Performance comparison of early action recognition on the SYSU 3DHOI Dataset. The same feature is used in all methods for fair comparisons. Refer to Figure 3(c) for more results.

Observation	Methods					
Ratios	[32]	[34]	ARN	EARN	EARN+GR	EARN+GR+TCE
0.3	51.08%	35.83%	33.39%	49.01%	54.68%	**56.06%**
0.5	62.14%	45.83%	54.64%	59.01%	63.24%	**64.35%**
0.8	69.32%	58.75%	67.67%	65.44%	68.61%	**69.94%**
Average	58.51%	45.58%	50.01%	55.81%	60.40%	**61.60%**

use the skeleton. We compared with [34] using their reported results based on skeleton. When the observation ratio is 0.1, the proposed EARN + GR + TCE achieves an accuracy of 41.11%. Compared to the performance of the methods in [32] and [34], the improvements are 10.1% and 15.9%, respectively.

4.6 Comparison with Pair-wise Distance

In this paper, we enforce the similarity between the full and partial sequences based on the mean and variance. When using the pair-wise distance between each pair of the full and the partial sequences as the regularizer to enhance their similarity, the prediction accuracy averaged across all observation ratios on the NTU Dataset is 38.91%, which is 10.5% worse than the proposed method. The partial sequences at the early stage are quite different from the full sequences. Some sequences at the beginning do not contain any motion, which introduces outliers of the samples. Using the pair-wise distance to enforce every pair of the full sequence and the partial sequence (including the outliers) to be similar makes the network difficult to converge, and results in poor performance. The proposed method minimizes the mismatch between the full and the partial sequences based on their statistical properties, which is capable of reducing the effects of the outliers and yields better performances.

4.7 Parameter Analysis

In this section, we evaluate the influence of λ in Equation 11 on the NTU Dataset. The prediction performances of the proposed method using different λ are shown in Table 5. When $\lambda = 0$, the global regularizer is not used in the network learning. As shown in Table 5, the prediction accuracies are improved when λ is increased. It clearly shows the advantage of the global regularizer for better early action recognition. Note that when λ is assigned a large value, the performance at the early stage is slightly worse. The reason might be that the network focuses on the learning of the global pattern of the full sequences, which leads to the under-fitting problem of the partial sequences at the early stage.

Table 5. Impact of the parameter λ on early action recognition performance on the NTU Dataset.

Observation			λ		
Ratios	0	0.1	0.5	1	2
0.3	27.87%	29.79%	33.80%	33.46%	33.04%
0.5	47.13%	51.92%	53.97%	54.28%	55.37%
0.8	59.88%	65.19%	65.39%	66.21%	67.49%
Average	43.77%	47.48%	49.10%	49.41%	50.08%

5 Conclusion

In this paper, we have proposed a new framework for skeleton-based early action recognition, which explicitly exploits the global information from the partial sequences using a global regularizer. The classification loss of the network is a temporal-aware cross-entropy, which prevents the network from over-fitting the partial sequences with small observation ratios. The proposed method has been tested on three benchmark datasets. Experimental results clearly show the advantages of the proposed method for early action recognition.

Acknowledgment

We greatly acknowledge NVIDIA for providing a Titan XP GPU for the experiments involved in this research. This work was partially supported by Australian Research Council grants DP150100294, DP150104251, and DE120102960.

References

1. Ryoo, M.S.: Human activity prediction: Early recognition of ongoing activities from streaming videos. In: ICCV, IEEE (2011) 1036–1043
2. Lan, T., Chen, T.C., Savarese, S.: A hierarchical representation for future action prediction. In: ECCV, Springer (2014) 689–704
3. Ma, Q., Shen, L., Chen, E., Tian, S., Wang, J., Cottrell, G.W.: Walking walking walking: Action recognition from action echoes. In: IJCAI, AAAI Press (2017) 2457–2463

4. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: CVPR, IEEE (2012) 1290–1297
5. Vemulapalli, R., Arrate, F., Chellappa, R.: Human action recognition by representing 3d skeletons as points in a lie group. In: CVPR, IEEE (2014) 588–595
6. Du, Y., Wang, W., Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition. In: CVPR, IEEE (2015) 1110–1118
7. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: NTU RGB+D: A large scale dataset for 3D human activity analysis. In: CVPR, IEEE (2016)
8. Zhu, W., Lan, C., Xing, J., Zeng, W., Li, Y., Shen, L., Xie, X., et al.: Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In: AAAI. Volume 2., AAAI Press (2016) 8
9. Liu, J., Shahroudy, A., Xu, D., Wang, G.: Spatio-temporal LSTM with trust gates for 3D human action recognition. In: ECCV, Springer (2016) 816–833
10. Ke, Q., Bennamoun, M., An, S., Boussaid, F., Sohel, F.: A new representation of skeleton sequences for 3D action recognition. In: CVPR, IEEE (2017)
11. Niebles, J.C., Chen, C.W., Fei-Fei, L.: Modeling temporal structure of decomposable motion segments for activity classification. In: ECCV, Springer (2010) 392–405
12. Wang, L., Qiao, Y., Tang, X.: Latent hierarchical model of temporal structure for complex activity classification. IEEE Transactions on Image Processing **23** (2014) 810–822
13. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: CVPR, IEEE (2015) 2625–2634
14. Kong, Y., Fu, Y.: Human action recognition and prediction: A survey. arXiv preprint arXiv:1806.11230 (2018)
15. Mahmud, T., Hasan, M., Roy-Chowdhury, A.K.: Joint prediction of activity labels and starting times in untrimmed videos. In: ICCV, IEEE (2017) 5784–5793
16. Bütepage, J., Black, M.J., Kragic, D., Kjellström, H.: Deep representation learning for human motion prediction and classification. In: CVPR, IEEE (2017) 2017
17. Ke, Q., Liu, J., Bennamoun, M., An, S., Sohel, F., Boussaid, F.: Chapter 5 - computer vision for humanmachine interaction. In Leo, M., Farinella, G.M., eds.: Computer Vision for Assistive Healthcare. Academic Press (2018) 127–145
18. Tang, C., Li, W., Wang, P., Wang, L.: Online human action recognition based on incremental learning of weighted covariance descriptors. Information Sciences **467** (2018) 219–237
19. Rahmani, H., Mahmood, A., Huynh, D., Mian, A.: Histogram of oriented principal components for cross-view action recognition. PAMI **38** (2016) 2430–2443
20. Rahmani, H., Mian, A., Shah, M.: Learning a deep model for human action recognition from novel viewpoints. PAMI (2018)
21. Rahmani, H., Bennamoun, M.: Learning action recognition model from depth and skeleton videos. In: ICCV, IEEE (2017)
22. Rahmani, H., Mian, A.: 3d action recognition from novel viewpoints. In: CVPR, IEEE (2016) 1506–1515
23. Wang, P., Li, W., Ogunbona, P., Wan, J., Escalera, S.: Rgb-d-based human motion recognition with deep learning: A survey. Computer Vision and Image Understanding (2018)
24. Ke, Q., An, S., Bennamoun, M., Sohel, F., Boussaid, F.: Skeletonnet: Mining deep part features for 3-d action recognition. IEEE Signal Processing Letters **24** (2017) 731–735

25. Ke, Q., Bennamoun, M., An, S., Sohel, F., Boussaid, F.: Learning clip representations for skeleton-based 3d action recognition. IEEE Transactions on Image Processing **27** (2018) 2842–2855
26. Liu, J., Wang, G., Hu, P., Duan, L.Y., Kot, A.C.: Global context-aware attention lstm networks for 3d action recognition. In: CVPR. Volume 7., IEEE (2017)
27. Ryoo, M.: Human activity prediction: Early recognition of ongoing activities from streaming videos. In: ICCV, IEEE (2011) 1036–1043
28. Kong, Y., Kit, D., Fu, Y.: A discriminative model with multiple temporal scales for action prediction. In: ECCV. (2014) 596–611
29. Ke, Q., Bennamoun, M., An, S., Boussaid, F., Sohel, F.: Human interaction prediction using deep temporal features. In: ECCVW, Springer (2016) 403–414
30. Ke, Q., Bennamoun, M., An, S., Sohel, F., Boussaid, F.: Leveraging structural context models and ranking score fusion for human interaction prediction. IEEE Transactions on Multimedia (2017)
31. Jain, A., Singh, A., Koppula, H.S., Soh, S., Saxena, A.: Recurrent neural networks for driver activity anticipation via sensory-fusion architecture. In: ICRA, IEEE (2016) 3118–3125
32. Aliakbarian, M.S., Saleh, F., Salzmann, M., Fernando, B., Petersson, L., Andersson, L.: Encouraging lstms to anticipate actions very early. In: ICCV, IEEE (2017)
33. Farha, Y.A., Richard, A., Gall, J.: When will you do what?-anticipating temporal occurrences of activities. arXiv preprint arXiv:1804.00892 (2018)
34. Hu, J.F., Zheng, W.S., Ma, L., Wang, G., Lai, J.: Real-time RGB-D activity prediction by soft regression. In: ECCV, Springer (2016) 280–296
35. Liu, J., Shahroudy, A., Wang, G., Duan, L.Y., Kot, A.C.: Ssnet: Scale selection network for online 3d action prediction. In: CVPR, IEEE (2018) 8349–8358
36. Herath, S., Harandi, M., Porikli, F.: Learning an invariant hilbert space for domain adaptation. arXiv preprint arXiv:1611.08350 (2016)
37. Hubert Tsai, Y.H., Yeh, Y.R., Frank Wang, Y.C.: Learning cross-domain landmarks for heterogeneous domain adaptation. In: CVPR, IEEE (2016) 5081–5090
38. Baktashmotlagh, M., Harandi, M., Salzmann, M.: Distribution-matching embedding for visual domain adaptation. The Journal of Machine Learning Research **17** (2016) 3760–3789
39. Pan, S.J., Tsang, I.W., Kwok, J.T., Yang, Q.: Domain adaptation via transfer component analysis. IEEE Transactions on Neural Networks **22** (2011) 199–210
40. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. (2012) 1097–1105
41. CMU: CMU graphics lab motion capture database. In: http://mocap.cs.cmu.edu/. (2013)
42. Hu, J.F., Zheng, W.S., Lai, J., Zhang, J.: Jointly learning heterogeneous features for RGB-D activity recognition. In: CVPR, IEEE (2015) 5344–5352