INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 20, Issue: 5, Month: October, Year: 2025
Article Number: 6882, https://doi.org/10.15837/ijccc.2025.5.6882

CCC Publications
AGORA UNIVERSITY PRESS

Recognizing A Complex Human Behaviour via A Shallow Neural Network with Zero Video Training Sample

Q. Xie, W. Lu, W. Yang, K. Xiong, L. Zhang, L. Yao

Qianglai Xie
the Center of Collaboration and Innovation
Jiangxi University of Technology
330098 Nanchang, Jiangxi, China
qianglai_xie@163.com

Wei Lu
the Center of Collaboration and Innovation
Jiangxi University of Technology
330098 Nanchang, Jiangxi, China
vluwei@163.com

Wei Yang
the Center of Collaboration and Innovation
Jiangxi University of Technology
330098 Nanchang, Jiangxi, China
yang.wei@163.com

Keyun Xiong
College of Computer Science
Jiangxi University of Chinese Medicine
330004 Nanchang, Jiangxi, China
20032023@jxutcm.edu.cn

Lei Zhang
Hanlin Hangyu (Tianjin) Industrial Co., Ltd.
301800 Tianjin, China
sanshilei@126.com

Leiyue Yao*
College of Computer Science
Jiangxi University of Chinese Medicine
330004 Nanchang, Jiangxi, China
*Corresponding author: leiyue_yao@163.com

Abstract

In contrast to human action recognition (HAR), understanding complex human behaviour (CHB), consisting of multiple basic actions, poses a significant challenge for researchers due to its extended duration, numerous types, and substantial data-labeling expenses. In this paper, a new approach to recognize CHB from a semantic point of view is proposed, which can be roughly summarized as judging by action quantization and action combination similarity. To fully evaluate the effectiveness of our method, the self-collected dataset – HanYue Action3D is extended to become the first public skeleton-based dataset with complex behavior samples and temporal calibration. Experimental results have demonstrated the feasibility and universal superiority of our method. Moreover, our method's zero-shot learning capability bridges the divide between laboratory settings and real-world applications.

Keywords: complex human behaviour recognition, temporal action localization, motion data structure, action encoding, skeleton-based action recognition

1 Introduction

Human motion recognition (HAR) has been the focus of computer vision research because of its broad application prospects[1]. Recent years, researchers have received significant achievements via utilizing the deep learning technology. Convolutional neural networks (CNNs), Long Short-Term Memory (LSTM) networks, Graph Convolutional Networks (GCNs), and their variants are widely used for recognizing human actions, continuously improving detection accuracy.

In the early stages of deep learning development, the most commonly used features for Human Activity Recognition (HAR) were hand-crafted, such as space-time volume features, Scale-invariant Feature Transform (SIFT), Histogram of Oriented Gradients (HOG), body contour and silhouette features, and motion trajectories[2]. The approaches relying on these hand-crafted features followed a common approach, which can be summarized as "interpretable numerical features + classifier"[3]-[6]. In this process, the first step involves carefully designing the features with hand-crafted attributes to represent human actions in numerical form. Once these actions were translated into numbers, classifiers such as the support vector machine (SVM) were trained on these quantified features and their labels, as evidenced by studies in human behavior detection, pedestrian detection, group anomaly detection, and student behavior diagnosis. Following the training phase, the classifier, leveraging the same quantification approach, can recognize new human actions, as demonstrated by the successful application of SVM algorithms in human behavior feature recognition.

While approaches based on hand-crafted features offer benefits in terms of interpretability and efficiency[7], it is important to recognize their inherent limitations, which lead to stagnation in accuracy and a lack of generalizability. Since 2012, deep learning has shown remarkable potential across numerous fields, including Human Activity Recognition (HAR), by effectively addressing the challenges that stemmed from the dependency on manually engineered features.

After AlexNet secured the top position in the ImageNet competition in 2012, 2D convolutional neural network (2D-cnn) has been widely recognized in various fields[8]-[9]. These 2D-CNN methods aim to extract deep features from pre-processed motion images, incorporating both spatial and temporal motion details. Techniques such as optical flow (OF) images [10]-[11], motion history images (MHI) [12], static history images (SHI) [13], motion energy images (MEI) [14], and various additional adaptations [15]-[22] are commonly employed to assist neural networks in learning discriminative features. Although methods based on 2D-CNNs exhibit high accuracy in short-duration human activity recognition (HAR) videos, they encounter difficulties with longer videos. In this context, 3D convolutional networks (3D-CNNs) [23]-[24], along with LSTM and its variants [25]-[28], provide a remedy. However, all these approaches are unable to utilize explicitly the structural topology of the joints, which is crucial for HAR. Given that graphs are ideally suited for modeling the human skeleton, graph convolutional networks (GCN) have driven the emergence of various GCN-based methods [29]-[32], significantly improving HAR accuracy. In the latest methods, researchers combined CNN and GCN to achieve better recognition accuracy.

Current methods for single action detection have reached accuracies of 96% or higher. Yet, the majority of researchers concentrate mainly on handling straightforward action recognition or employ comparable methods in addressing CHBs recognition. While it is theoretically possible to recognize CHBs in the same manner as HAR (Human Activity Recognition), doing so presents significant challenges in practice due to the following two limitations.

(1)A CHB consists of a sequence of fundamental human actions, resulting in a duration that significantly exceeds that of a single action. In addition to the challenges associated with collecting and labeling CHB samples, training deep neural networks (DNNs) poses a considerable challenge given the current capacity of GPUs.

(2)Unlike basic human actions, CHBs are not enumerable. Although numerous transformer-based solutions exist, accurately recognizing a new type of CHB that was not present during the training phase remains impossible.

Therefore, it is important to explore innovative methods for recognizing CHBs in the current technological background. Human language provides valuable insights, as CHBs share many similarities with it. Just as a finite vocabulary can generate an endless array of sentences and articles through diverse arrangements and combinations, so too can a limited set of basic actions produce an infinite variety of CHBs. The CHB can be inferred and recognized by combinations from a semantic point of view if we consider basic human behaviors as words and the CHB as sentences or articles. Along this thought, existing temporal action localization (TAL) techniques are particularly suitable for detecting action sequences in CHBs.

The goal of temporal action localization (TAL) is to concurrently classify actions while identifying the start key frame (SKF) as well as end key frame (EKF) for each action instance within an unedited video. Generally, TAL seeks to address two main issues: 1) When should the actions begin and end? and 2) Which categories should these actions fall? [33]

In previous studies, researchers have employed a two-stage pipeline, known as 'propose and classify', to accomplish the task of action recognition. In this two-stage framework, it firstly generates several temporal proposals, followed by classification into various action classes. Then, the temporal boundary regressions will be applied to these proposals. Considering the recent progress in action recognition and boundary regression techniques, the primary hurdle faced by two-stage methodologies lies in the creation of temporal proposals.

The most commonly used mechanisms include sliding windows [34] and anchor-based strategies [35]-[37]. While these two-stage methods have demonstrated impressive performance, two major shortcomings that cannot be ignored: 1) The setting of anchors is a complex task because these approaches need to manually configure the anchors and their scales, restricting flexibility; 2) Because they rely on temporal annotations, these methods are often expensive and time-consuming, which significantly hinders their application in real-world scenarios.

As opposed to the two-stage approach, the one-stage approach [38]-[40] is also referred to as the anchor-free methods, which can handle both classification as well as proposal generation at the same time and can train easily in an end-to-end manner. Essentially, the one-stage approach adheres to the basic principles of the two-stage approach but differs by necessitating only video-level labeling for frame-level action estimation. The one-stage approach generates proposals that are more flexible in length than the fixed-length temporal proposals used in the two-stage approaches. Nevertheless, the precision of one-stage methods tends to surpass that of two-stage methods.

This paper proposes a novel approach for CHB recognition that does not require training samples, offering a generalized solution for various CHB detections. Our proposed approach integrates a 2D-CNN-based methodology with a one-stage TAL technique. Firstly, drawing upon the description that 'human behavior consists of a series of actions', CHB is discretely represented as a sequence of simple actions, with each action further defined by an action word and its corresponding time duration. Using this definition, the problem of CHB recognition can be redefined into a semantic similarity judgment problem. Therefore, a set of pre-existing word vectors can be used to construct a model for processing the input data. Lastly, training samples can be readily generated at the semantic level, thereby obviating the need for collecting and labeling video samples.

The paper is organized as follows: An overview of current state-of-the-art TAL methods and approaches is given in Section 2. Section 3 describes our HAR model and TAL algorithm in detail. Section 4 details the experimental procedures and findings. Finally, Section 5 summarizes the paper and discusses potential work for the future.

2 Related work

Compared with basic human actions, CHBs possess greater significance. Understanding the semantic information included in CHBs is helpful to develop advanced human-machine interaction applications. Nevertheless, the prevailing method for identifying CHBs remains the rule-based process of 'sample labeling, model training, and CHB classification'. This approach, while technically feasible, will inevitably result in massive manual labeling works and a strong reliance on computing power. In our proposed method, the critical step involves the precise extraction of fundamental human actions and their corresponding features from CHB videos, ensuring the correct sequence is maintained. To achieve our CHB recognition objective, we leverage a lightweight network structure for skeleton-based HAR, as demonstrated by the Double-feature Double-motion Network (DD-Net), in conjunction with one-stage temporal action localization (TAL) techniques.

2.1 Skeleton-based HAR methods

Recent research in HAR has prominently featured the integration of depth images with deep learning technology. The problem of deformation caused by various human physiques can be effectively solved by extracting the human skeleton from depth information, so as to achieve the purpose of efficiently extracting relevant information [41]. Y. Du et al. [42] transformed human actions into color images and employed a convolutional neural network (CNN) for classification. The three-dimensional coordinates (x, y, z) of the joints were creatively stored in the three color channels (R, G, B) of the pixel. In [43], the action image from reference [39] was enhanced using a tree structure skeleton image (TSSI), which preserves the spatial relationships among joints more effectively, thereby improving accuracy. In our previous works [44]-[45], a custom data structure called Dense Joint Motion Matrix (DJMM) is introduced based on these method to store joint motion features with high accuracy. This foundation enabled us to propose an effective data augmentation technique, a temporal scale unifying strategy,and an updated multi-scale ZF-Net to achieve high-accuracy HAR. The skeleton-based HAR approach mentioned in this paper exhibits superb accuracy in experiments, thus, we continue to utilize it in this paper to detect basic human actions. Further details will be elaborated in Section 3.1.

2.2 One-stage TAL

According to reference [33], high-quality temporal action proposals ought to exhibit three key attributes:

1)Flexible temporal length.

2)Precise temporal boundaries.

3)Reliable confidence scores.

A bottom-up method is used in the one-stage TAL method to identify actions by analyzing frame-level action scores. For instance, the BSN method [46] first generates start and end probabilities. It then uses a proposal generation module to derive proposals, and the proposal evaluation module obtains confidence scores finally. Lin et al. [47] enhanced the BSN method into an end-to-end framework using a BM-Layer. In [48], a different end-to-end approch called DBG was mentioned by this paper, in which the action-aware completeness regression module creates the completeness map as well as the temporal boundary classification module provides the boundary map. This enables proposals to be generated by combining the two maps in an effective way. In [49], this article covers a more accurate method known as BSN++, it offers accurate the boundary map as well as confidence map via a proposal relation block and a complementary boundary generator.

Despite the emergence of new TAL methods, many of them still struggle to recognize basic human actions in extended behavioral videos with brief segments. To address the challenge of enhancing the precision of TAL, we previously introduced a one-stage method that leverages the Dense Joint Motion Matrix (DJMM). This method is specifically designed to detect human actions within behavioral videos, as detailed in our research. This method remains in use for extracting information on fundamental human actions. Section 3.2 provides further details.

3 Proposed approach

A CHB comprises multiple basic actions arranged in chronological order. Following this approach, recognizing CHBs without the need for training samples is theoretically achievable by analyzing these action combinations' semantic similarity. The proposed method incorporates three primary technologies: action recognition, temporal action localization, and semantic similarity judgment. Figure 1 presents an overview of the method, which is broken down into three main steps.

Step 1: Identify the basic actions. The CHB video is firstly segmented into clips with consistent frame counts. A multi-scale CNN is then applied to identify the actions in each clip. This step has two key components: First, human action videos are characterized by 26 joint motion features represented within a floating-point matrix. Second, the float matrix is analyzed via using a multi-scale CNN model. The output which from the CNN model includes about four key elements: the start key frame (SKF), the end key frame (EKF), the action classification (action), the and confidence (C).

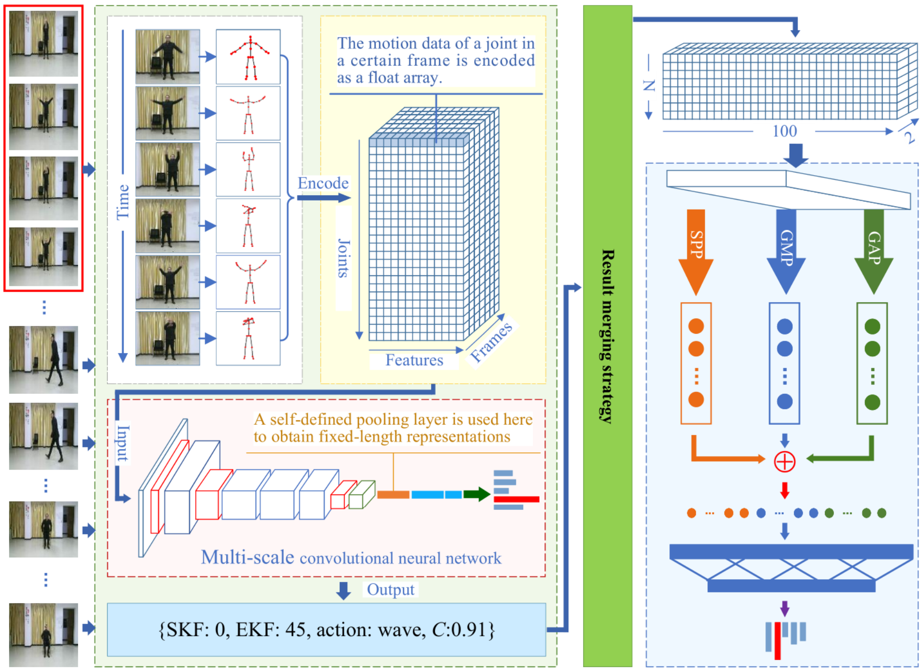

Time
The motion data of a joint in a certain frame is encoded as a float array.
Encode
Joints
Features
Frames
Input
A self-defined pooling layer is used here to obtain fixed-length representations
Multi-scale convolutional neural network
Output
{SKF: 0, EKF: 45, action: wave, C:0.91}
Result merging strategy
N
100
2
SPP
GMP
GAP

Figure 1: The general block of our proposed method.

Step 2: Combine the results from Step 1. The CNN model produces a series of four-tuple outputs, each spanning the same duration from SKF to EKF. The goal at this step is to merge these outputs to precisely identify the SKF and EKF for each basic action in the video.

Step 3: Recognize the CHB by analyzing the semantic similarity between action combinations. During this phase, the word vectors are utilized to illustrate action words. This allows the basic action along with its duration to be depicted as a float matrix of 3D, which can then be processed and identified by a neural network.

The subsequent sections will provide comprehensive details on the algorithms and the neural network structures utilized at each stage.

3.1 A CNN-based action recognition method

Our previous research demonstrated that integrating CNN models with human skeleton motion data is an effective strategy for HAR, as evidenced by the growing body of literature in this field. The action representation in the proposed method builds on this previous work. Because the human body is described as a system of articulated joints and rigid bones, human actions are depicted as skeletal movements [1]. In this method, Motion features of the 26 joints, including speed, displacement, and direction, are computed frame by frame to represent human actions.

Fig. 2 illustrates a 3D float matrix (DJMM) used to store joint motion features, with the x, y, and z axes representing motion features, joints, and frames, respectively. Let j_m^n represent the m^{th} joint of n^{th} frames in a clip, then $F\left(j_m^n, i\right)$ indicates the i^{th} feature of j_m^n, thus, the quantified human motion in a clip,σ , can be shown as:

$$\sigma = \left[\begin{bmatrix} F\left(j_0^0, 0\right) & \cdots & F\left(j_0^0, i\right) \\ \vdots & \ddots & \vdots \\ F\left(j_{24}^0, 0\right) & \cdots & F\left(j_{24}^0, i\right) \end{bmatrix}, \ldots, \begin{bmatrix} F\left(j_0^n, 0\right) & \cdots & F\left(j_0^n, i\right) \\ \vdots & \ddots & \vdots \\ F\left(j_{24}^n, 0\right) & \cdots & F\left(j_{24}^n, i\right) \end{bmatrix} \right] \tag{1}$$

It is crucial to recognize that the structure's adaptability stems from the adjustable lengths of the x and z axes. Motion features can be altered, including being added or removed, based on the required balance between accuracy and efficiency. Furthermore, various action durations can be configured to enhance prediction confidence.

To handle the variable-sized 3D float matrix, especially the time dimension variability, we previously introduced a CNN with multi-scale reshaping strategy using a spatial pyramid pooling (SPP)

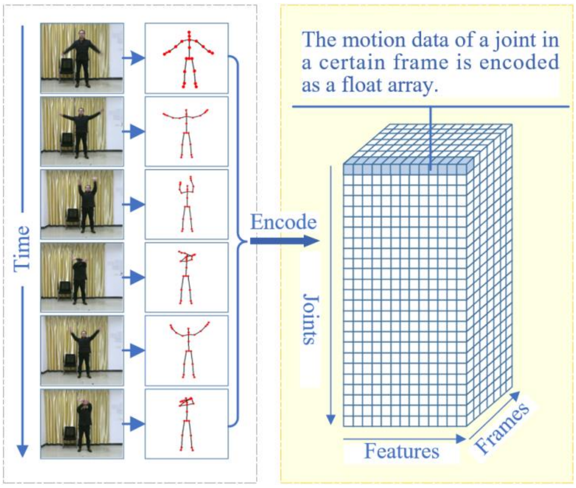

Time
Encode
The motion data of a joint in a certain frame is encoded as a float array.
Joints
Features
Frames

Figure 2: The 3D float matrix which is used for storing the 26 joints' motion data of an action.

layer. In this study, we further refined the technique by incorporating the SPP layer with GAP and GMP layers, forming the novel GMS layer. In comparison to our previous method, the integration of an adaptive GAP layer has significantly enhanced the model's accuracy by preserving global features from the hidden layers, while the GMP layer concentrates on capturing the unique features. **Fig.3** demonstrates the multi-scale architecture.

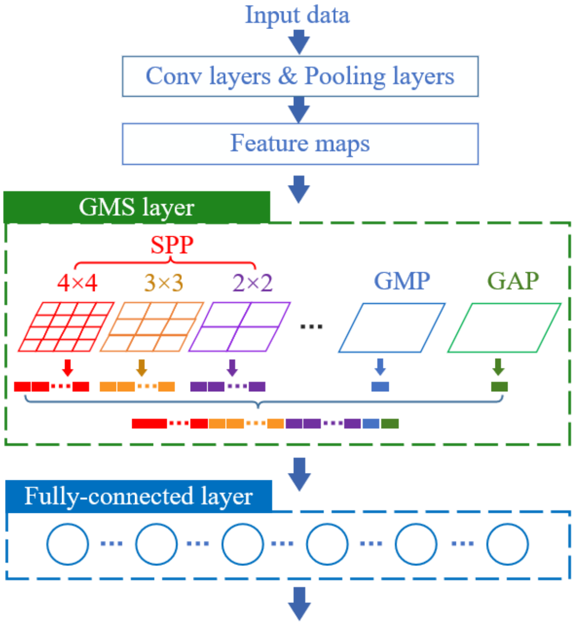


Figure 3: The structure of the GMS layer. It contains 1 SPP layer, 1 GMP layer and 1GAP layer. The SPP layer uses a pooling parameter of [4, 3, 2], which resulting in a fixed output length of 31 dimensions.

The GMS layer's final output size is determined using Equation (2).

$$L = \sum_{i}^{n} N_f \times P_i \tag{2}$$

where L stands for the GMS layer's output length. N_f represents the count of feature maps, and $P = (P_0, P_1, \ldots, P_i)$ denotes the pooling layers, which used in the SPP layer. For example, with $P =$ [(4,4), (3, 3), (2, 2)]and 128 feature maps before the SPP layer, the output size of the GMS layer is calculated as $(4 \times 4 + 3 \times 3 + 2 \times 2) \times 128 + 128 + 128 = 3968$.

3.2 The effectiveness of the TAL method

The primary challenge in our proposed approach lies in accurately identifying the start key frame (SKF) and end key frame (EKF) of basic actions. In our latest research work [45], we presented an effective and efficient one-stage TAL method designed to extract the SKFs and EKFs of actions. This paper introduced a consolidation approach, namely the large-scale-first window merging strategy, which is designed to effectively combine detected results by prioritizing the merging of larger datasets. However, it still encounters a minor limitation when handling very short-duration actions, which can negatively impact the detection of complex human behaviors (CHB). **Fig.4** provides an illustrative example of an unsatisfactory TAL outcome.

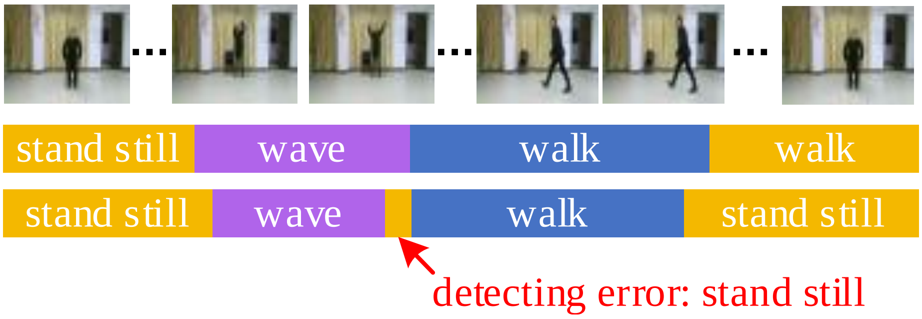


Figure 4: An example of the trivial action deficiency.

Since predictions based on small-duration actions lack meaning and are likely to detect errors with high probability, they can be eliminated through a specific strategy. In the proposed method, these small-duration actions were merged into the previous action.

The pseudo-code below shows the improved large-scale-first window merging strategy. Combined with the action recognition model, it further improves the precision of our previous work [45].

3.3 Zero-shot based approach for CHB recognition

3.3.1 the theoretical basis of the method

Consider an example of complex human behavior: "Alex is excitedly walking back and forth, while waving his hand." Based on the CHB definition encompassing multiple basic actions, this example can be discretely represented as [[walk, wave], [wave, walk], [walk, wave, wave, walk], ...]. In theory, additional action features like duration, direction, etc., are incorporated to describe a CHB more precisely. Moreover, word vectors are highly useful for uncovering semantic meanings. The semantics of different word combinations can be easily calculated. An example of using word combinations to discover meaning is shown in **Fig. 5**.

From **Fig.5**, the newly calculated word vector, formed by combining "computer," "basketball," and "simulation," show the shortest distance from the word vector for "game."

This indicates that the vector for the new word bears a meaning akin to "game," thus justifying its classification within the same category.

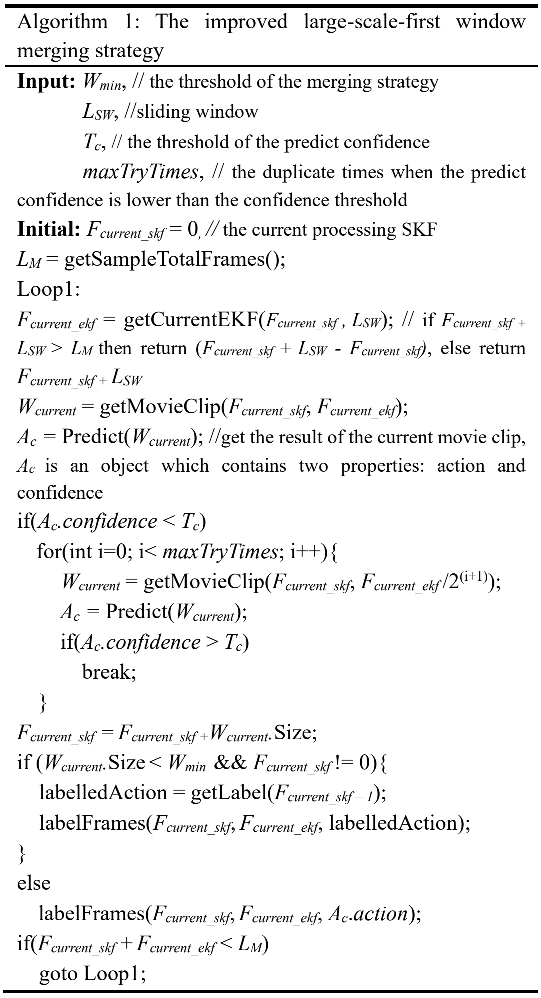
The transformative power of word vectors in revealing semantic meaning has been pivotal to our ground breaking innovations.

From one perspective, this feature enables us to train a CNN model without relying on video samples. Human language, as we know, is capable of describing highly complex concepts, and human behavior is undoubtedly included in this. These behaviors can be succinctly described using a few key terms. Let's assume that a basic human action can be expressed mathematically as follows:

$$\xi = \{n, A\} \tag{3}$$

where n represents the action word, while A denotes the action's features. Based on this, a complex human behavior is shown mathematically as follows:

$$\sigma = \left(\left[\xi_1, \xi_2, \ldots \right], \left[\xi_m, \xi_n, \ldots \right], \left[\xi_x, \xi_y, \ldots \right], \ldots \right) \tag{4}$$



Algorithm 1: The improved large-scale-first window merging strategy

```
Input: $W_{min}$, // the threshold of the merging strategy
       $L_{SW}$, //sliding window
       $T_c$, // the threshold of the predict confidence
       $maxTryTimes$, // the duplicate times when the predict confidence is lower than the confidence threshold
Initial: $F_{current\_skf} = 0$, // the current processing SKF
$L_M$ = getSampleTotalFrames();
Loop1:
$F_{current\_ekf}$ = getCurrentEKF($F_{current\_skf}$ , $L_{SW}$); // if $F_{current\_skf} + L_{SW} > L_M$ then return ($F_{current\_skf} + L_{SW}$ - $F_{current\_skf}$), else return $F_{current\_skf} + L_{SW}$
$W_{current}$ = getMovieClip($F_{current\_skf}$, $F_{current\_ekf}$);
$A_c$ = Predict($W_{current}$); //get the result of the current movie clip, $A_c$ is an object which contains two properties: action and confidence
if($A_c.confidence < T_c$)
    for(int i=0; i< $maxTryTimes$; i++){
        $W_{current}$ = getMovieClip($F_{current\_skf}$, $F_{current\_ekf}/2^{(i+1)}$);
        $A_c$ = Predict($W_{current}$);
        if($A_c.confidence > T_c$)
            break;
    }
$F_{current\_skf} = F_{current\_skf} + W_{current}$.Size;
if ($W_{current}$.Size< $W_{min}$ && $F_{current\_skf}$ != 0){
    labelledAction = getLabel($F_{current\_skf - 1}$);
    labelFrames($F_{current\_skf}$, $F_{current\_ekf}$, labelledAction);
}
else
    labelFrames($F_{current\_skf}$, $F_{current\_ekf}$, $A_c.action$);
if($F_{current\_skf} + F_{current\_ekf} < L_M$)
    goto Loop1;
```

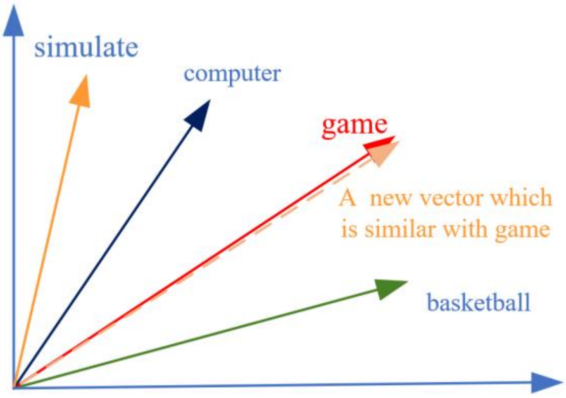


Figure 5: The illustration of semantic discovery by combining several word-vectors with different meaning.

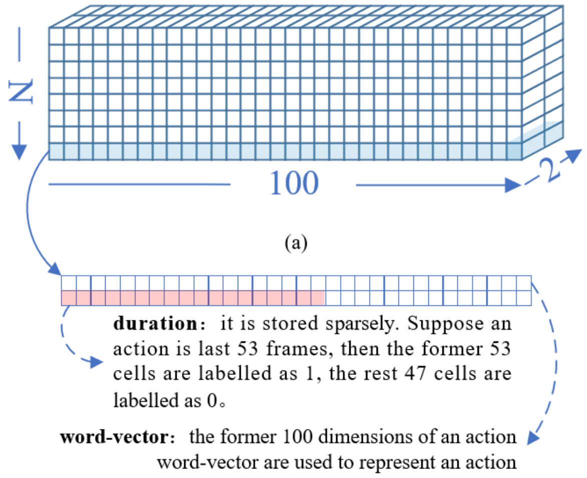
where σ stands for the CHB, $[\xi_1, \xi_2, ...]$,$[\xi_m, \xi_n, ...]$ and $[\xi_x, \xi_y, ...]$ represent the various basic action combinations of the complex human behavior.

Section 3.2 describes a method for extracting human actions and their durations related to the actions within a video. This enables the representation of a complex human behavior video as a series of integrated actions. Similarly, both equations (1) and (2) depict complex human behavior as a combination of actions. As a result, during both the training and inference stages of the DNN, we can utilize word-vectors for action words instead of relying on video samples.

Additionally, numerous pre-trained word-vectors are readily available for use. Even though these word-vectors are derived from various language models and different datasets, they share a similar structure — a one-dimensional floating-point array. Consequently, regardless of the word-vector model we choose, it is unnecessary to modify the architecture of DNN. Section 4 presents comprehensive experiments conducted to validate the versatility of the proposed approach.

3.3.2 The data structure

To evaluate the semantic similarity among various action combinations using a DNN, digitally representing action words is crucial. Word2Vec serves as the most effective solution for this task. Word-vectors typically come in two forms: one-hot encoding and distributed word representation (DWR). Because of dimensionality explosion issues, DWR has emerged as the primary representation in natural language processing (NLP). Numerous word-vectors are available, creatested using various language models and datasets, with most of them having dimensions that span from 100 to 300. Typically, the initial 100 dimensions suffice for distinguishing the meanings among different words. According to the definition given in equation (2), the CHB can be represented as a three-dimensional float matrix. The structure is illustrated in **Fig.6**.



duration: it is stored sparsely. Suppose an action is last 53 frames, then the former 53 cells are labelled as 1, the rest 47 cells are labelled as 0。

word-vector: the former 100 dimensions of an action word-vector are used to represent an action

Figure 6: The data structure of a CHB. (a) is a CHB's whole structure. (b) is the structure of a quantified action.

As illustrated in Fig. 6 (a), we assume that a CHB comprises N basic actions arranged sequentially in time, with each action being represented by the initial 100 dimensions of its word-vector, accompanied by its duration in frames. Consequently, the CHB is represented as a float matrix which size $N \times 100 \times 2$.

Notably, we employ the sparse method to store the duration of each action, which is shown in Fig. 6 (b). While a single numerical value (which has only one-dimensional) could suffice to capture the action's duration, this approach risks diminishing the significance of the duration feature during convolutional computations.

Another aspect can be considered is our approach utilizes two dimensions—one for the word-vector and another one is for the action duration. These two dimensions can represent the fundamental actions of humans. However, the inclusion of additional dimensions (or features) could improve accuracy or aid in distinguishing between similar human behaviors.

3.3.3 The neuro-network

As mentioned in section 3.3.2, intricate human behavior is shown by a 3D float matrix. The 100-dimensional action word vector is obtained from a well-trained word2vec model, and the action duration is stored using a sparse method. For example, a CHB consists of three fundamental actions, resulting in a matrix size of just 3 × 100 × 2, where the values in each cell of the dimensions contain significant information. With this understanding, we undertook an ambitious endeavor to develop the neural network. **Fig. 7** illustrates the structure. The main innovations can be summarized in three key points.

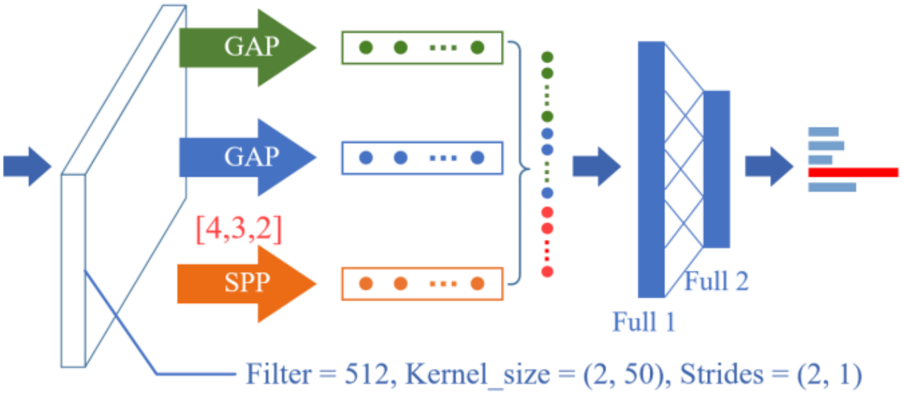


Figure 7: The structure the CHB recognition neuro network.

(1)As the highly abstract nature of the input data, the DNN model with a shallow architecture has been developed, leading to significant efficiency during both the training and testing phases.

(2)Deep features are extracted using a large convolutional kernel.

(3)A custom GMS layer replaces the Flatten layer to obtain an output of fixed length before the softmax layer. Composed of a GAP layer, a GMP layer, and an SPP layer, the GMS layer serves a crucial role.

4 Experiments and evaluation

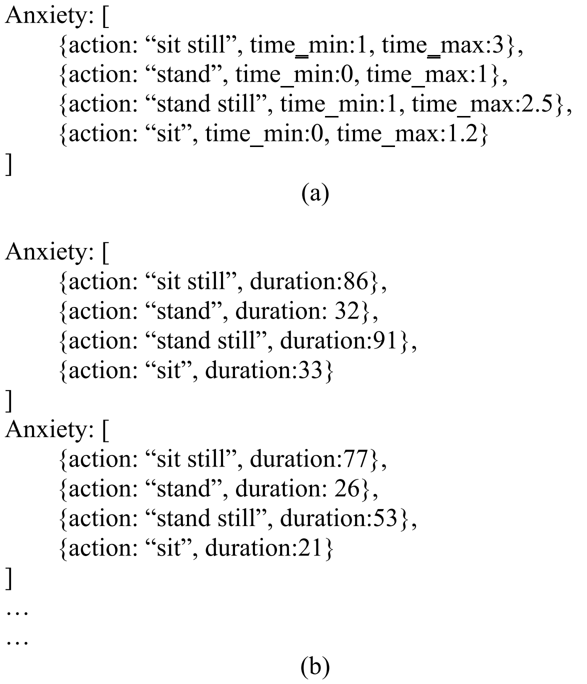
4.1 Implement details

This approach utilized TensorFlow 2.3 with GPU support and Keras for implementation. Our experiments were conducted on a desktop computer equipped with an Intel Core i7-8700K processor running at 3.70 GHz, an Nvidia GTX-1080ti GPU, and 64 GB of RAM operating at 3200 MHz.

To evaluate our proposed method, we utilized a dataset that we collected ourselves, named the HanYue-3D dataset, acquired with a Kinect v2.0 camera. This dataset comprises 15 types of simple actions: making a phone call, drinking, waving hands, checking a watch, brushing dust off clothing, falling, pushing a chair, jumping in place, standing up, remaining still, clapping while standing, walking, sitting, sitting still, and clapping while sitting. Nine subjects were instructed to perform these 15 activities three or four times. The Kinect v2.0 sensor recorded the locations of all 25 joints in a 3D coordinate system. In total, there are 413 samples, with each action type containing between 35 and 37 samples. Furthermore, for the purpose of action temporal localization, four types of complex behaviors were collected. Every complex behavior is made up of multiple simple actions that were previously described. The four complex behaviors are "sit still -> stand up -> brush dust off clothing" , "jump -> remain still -> jump -> wave hands","walk -> remain still -> jump -> wave hands -> sit down", and "clap while sitting -> stand up -> clap while standing -> jump -> wave hands". The HanYue-3D dataset is now available at: http://116.62.233.186:8878/HanYue-Action3D.zip.

4.2 The design of experiment

To thoroughly evaluate the effectiveness, efficiency, and flexibility of our proposed method in the absence of video training samples, the experiment focuses on three key aspects: the construction of



```
Anxiety: [
        {action: “sit still”, time_min:1, time_max:3},
        {action: “stand”, time_min:0, time_max:1},
        {action: “stand still”, time_min:1, time_max:2.5},
        {action: “sit”, time_min:0, time_max:1.2}
]
                                        (a)

Anxiety: [
        {action: “sit still”, duration:86},
        {action: “stand”, duration: 32},
        {action: “stand still”, duration:91},
        {action: “sit”, duration:33}
]
Anxiety: [
        {action: “sit still”, duration:77},
        {action: “stand”, duration: 26},
        {action: “stand still”, duration:53},
        {action: “sit”, duration:21}
]
…
…
                                        (b)
```

Figure 8: The definition and representation of the CHBs. (a) is the definition of a CHB, where the duration is a time range which is defined by "time_min" and "time_max". (b) shows the representations of a CHB, which are generated from its definition.

the semantic dataset, the selection of the language model and word vectors, and the design of the evaluation strategy.

4.2.1 The construction of semantic dataset

The primary motivation behind our approach is to recognize complex human behaviors (CHBs) without the need for video training samples due to the high costs associated with the collection and labeling process. Additionally, a CHB is supposed to be composed of various basic human actions. For example, the state of "anxiety" might be manifested as "someone repeatedly sitting and rising," or alternatively, as "someone continually checking their watch while pacing back and forth." Moreover, many action combinations exist that can describe "anxiety." Therefore, gathering sufficient samples to train a DNN model effectively is challenging, and in some cases, even impossible. However, defining a CHB semantically is significantly simpler. Taking "anxiety" as an example once more, it can be depicted as [sit, stand, sit, stand, ...] or [glance at the watch, walk, glance at the watch, walk, ...]. based on the equation(2) that was shown in Section 3.3.1, we established 8 CHBs and produced 248,000 samples using a program. **Fig. 8** illustrates the definition and generation strategy for the CHB sample, using "anxiety" as an example.

Fig.8(a) illustrates the strategy for defining complex human behaviors (CHBs). This definition follows a JSON format that consists of three properties. The properties "time_min" and "time_max" indicate the range of action duration. For example, action: "sit still", time_min: 1, time_max: 3 signifies sitting still for a duration ranging from 1 to 3 seconds. During the sample generation phase, the duration of each action is assigned a random value between 30 and 90 frames, as 1 second corresponds to 30 frames. **Fig.8(b)** displays two of the generated samples. Due to the structured nature of the CHB definition, it is straightforward to generate a significant quantity of samples programmatically in a very short time.

The advantages are twofold. First, high computational cost video samples can be substituted with word definitions (as illustrated in **Fig. 8**, leading to a significant reduction in computational costs. Second, once defined by words, a substantial number of CHB samples can be generated programmatically within a samll frames. Following this strategy, we defined 8 CHBs and produced 248,000 samples for the HanYue-3D dataset.

4.2.2 The selection of language model and word-vector

To obtain optimal accuracy for the CHB model, it is essential to choose widely-used language models and linguistic datasets and conduct a thorough comparison to determine the most suitable word vector for our proposed model. To enhance the model's generalization, word vectors with similar semantic meanings are crafted to be interchangeable, as demonstrated by the effectiveness of pre-trained word vectors in NLP tasks. For example, if the action word vector 'hit' is substituted with the word vector 'punch', our model should be expected to yield the same output.

4.2.3 The strategy of evaluation

(1)Validation Evaluation Strategy: Recognizing a complex human behavior (CHB) using a semantic similarity judging model is the final step of our work. The effectiveness of this step is crucial to the overall feasibility of the project. Therefore, the semantic similarity judging model have to be evaluated independently and evaluated thoroughly for validation. In our experiments, 80% of the samples generated according to the rules in section 4.2.1 are designated as the training samples, whereas the remaining 20% serve as the testing samples. If the accuracy exceeds 90%, the model can be deemed validated; otherwise, it is considered unsuccessful.

(2)Super-Parameter Combinations Selection Strategy: Due to the non-interpretability of deep neural networks, extensive experimentation is necessary at this stage. The combination of super-parameters that yields the best accuracy and efficiency will be selected as the optimal combination.

(3)Generalization Ability Evaluation Strategy: As noted in section 4.2.2, the CHB's generalization ability in semantic similarity assessment model is of significant importance, as there are numerous classic word vectors, and different words can have similar or even identical meanings. It is essential to verify the model's effectiveness with various word vectors. Moreover, demonstrating the model's validity when substituting the action keyword with another word with similar meaning is vital. a similar semantic meaning. In this experimental phase, all mainstream word vectors should be included. Additionally, the action key words defined in a complex human behavior should be replaced with other action words of similar meaning. If the model produces the same output in both scenarios, it is considered feasible; otherwise, it fails.

(4)Zero-Shot Capacity Evaluation Strategy: This represents the final evaluation of our approach. As shown in Fig. 1, the model's accuracy is closely tied to the human action recognition model, the temporal action localization model, as well as the CHB semantic similarity judging model; the results from all three models significantly influence the final accuracy. In this stage, experiments will be conducted on various CHBs that encompass several basic human actions included in the HanYue-3D dataset. Furthermore, to further assess the model's capacity, the CHBs should be performed by different volunteers who enact the basic human actions.

4.3 Experiment result and analysis

According to the experiment design in section 4.2.3, we conduct 3 types of experiments as follows.

4.3.1 Validation evaluation and Super-parameter combinations selection

The CHB semantic similarity judging model serves as the final step of our proposed method. The effectiveness of our method hinges on the validation of this model. Therefore, the validation evaluation experiments for the CHB semantic similarity judging model take precedence over other experiments. The initialization for this section of the experiments is established as follows:

Language model: word2vec;

Linguistic dataset: wiki;

The dimensions used of the word-vector: 100;

The neuro network: Refer to Fig.7;

The CHB semantic dataset: HanYue-3D semantic samples which are generated following the rules of Fig.8(b).

The super-parameters:

filters = (32, 64, 128, 256, 512)
kernel_size = (50, 75, 100)
activations = ("selu", "sigmoid", "tanh", "relu")
epochs = (5, 10, 20)
batch_sizes = (64, 128, 256, 512)

Based on the super-parameters mentioned above, a total of 720 combinations were tested. The outcomes from all 720 combinations lead to the following conclusions:

(1)The CHB semantic similarity judging model performed effectively across all super-parameter combinations, indicating that the input data structure defined in Fig. 6 integrates seamlessly with the model.

(2)The model achieved satisfactory training within 10 epochs across all 720 combinations due to the shallow architecture of the network.

(3)Among the activation functions tested, the sigmoid function exhibited the poorest performance. Therefore, in subsequent experiments, the "sigmoid" function will be excluded. **Fig.9** presents the comparisons among sigmoid, tanh, and relu..

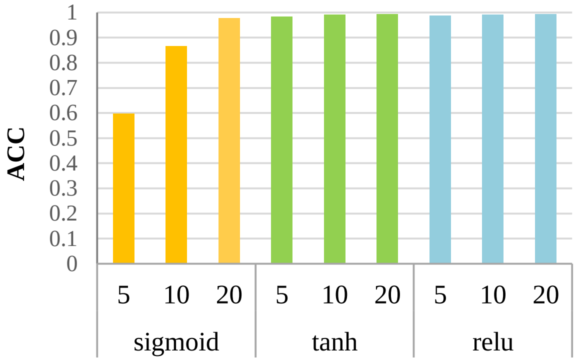


Figure 9: The comparison among different activation functions.

(4)The value of "batch_size" is not necessarily better when larger. If "batch_size" is set too high, the model will require more epochs for effective training.

(5)The values of "Filters" have a minimal impact on the model. This is because the word-vector is highly abstracted through the language model, and the sparse storage strategy diminishes the model's non-linear requirements. Nevertheless, theoretically, a rise in the quantity of feature maps can result in higher accuracy. Therefore, when initializing filters, this paper takes both overall efficiency and accuracy into account.

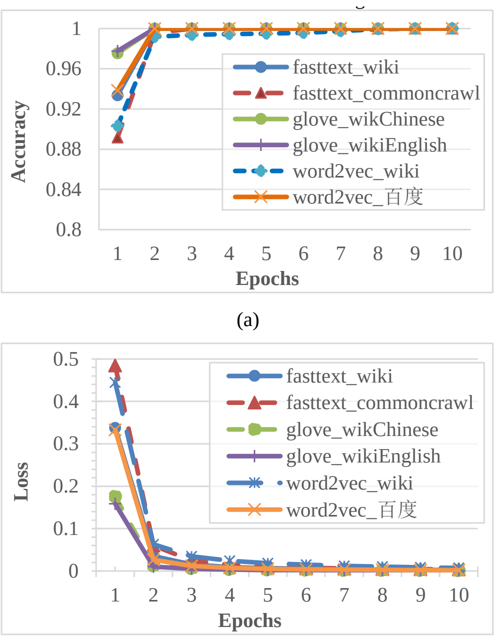
Based on the five conclusions above, the super-parameters of the neural network in Fig.7 are ultimately determined as follows:

filters = 128
kernel_size = (2, 50)
strides = (2, 1)
activation = "selu"
epochs = 10
batch_size =128

4.3.2 Generalization ability evaluation

A language model can be developed using various linguistic datasets, resulting in distinct word-vectors. To validate the model's performance across these different word-vectors, we selected six word-vectors derived from three common language models (FastText, GloVe, and Word2Vec) and two linguistic datasets for our experiments. The outcomes are illustrated in **Fig. 10**.

The results indicate that the model performed effectively with all six word-vectors. The neural network successfully completed training within 10 epochs, highlighting the model's efficiency. Moreover, the model can be readily applied without extensive training when a new CHB requires detection.



(a)

(b)

Figure 10: The TAL experiment results tested on HanYue-3D. The fine-tuned ZFNet was adopted as the backbone of the framework.

In accordance with the evaluation strategy outlined in section 4.3.2(3), we substituted several action words with others of similar meanings, such as “stand” replaced by “stand-up,” “wave” by “brandish,” “applaud” by “handclap,” and “applaud” by “clap.” The model continued to perform well, achieving an accuracy of 89.57%.

This segment of the experiments confirms that it is feasible to define the same CHB using different action words for the CHB semantic similarity judging model, as the Euclidean distance between two distinct words with similar meanings remains small. Consequently, it can be inferred that sentence-vectors can also effectively define a CHB. Additionally, since a sentence can encapsulate more nuances of a CHB, it is likely that sentence-vectors will be more efficient than word-vectors in differentiating similar CHBs.

4.3.3 Zero-shot capacity evaluation

The ultimate goal of our method is to recognize a CHB without relying on any video training samples. To assess its zero-shot capacity, the experiments are carried out according to the following criteria:

(1)The HAR model was trained using the complete set of basic human action video samples from the HanYue-3D dataset.

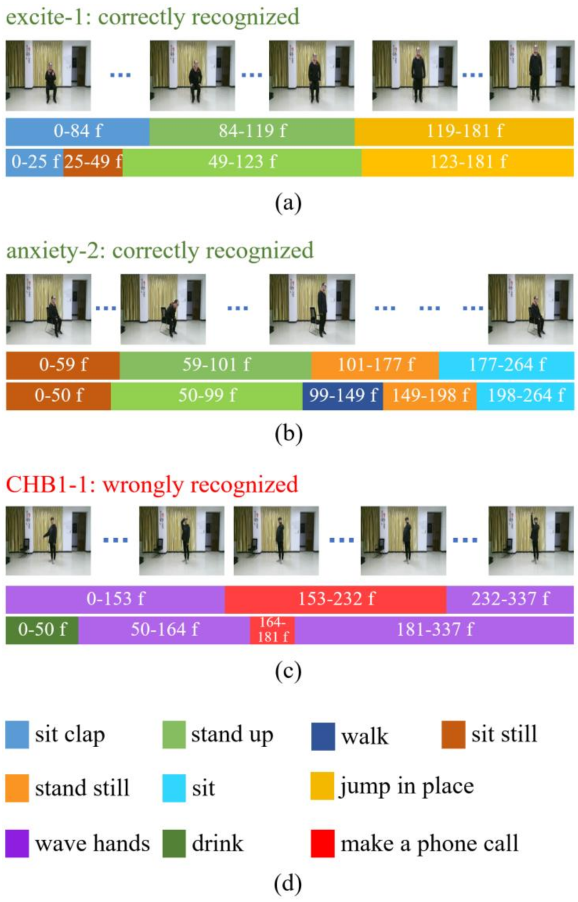
(2)Seven CHBs were selected as testing samples, which include anxiety, hover, excite, and four additional long-duration behaviors labeled as complex human behavior 1-4. These seven behaviors were performed by volunteers who were entirely different from those who demonstrated the basic human actions.

(3)Each of the seven CHBs was performed twice, resulting in a total of 14 CHB samples for assessment. Collecting additional number of test samples could enhance the precision of the evaluation, but the current samples are adequate to assess the effectiveness of the proposed method.

(4)Prior to application, the CHB semantic similarity assessment model was trained by using the semantic samples.

During our experiments, 10 CHB samples were correctly identified and 4 were incorrectly identified. Even though the accuracy was just 71.43%, this still demonstrates that the proposal model is functional. Furthermore, two key points are supposed to be emphasized: 1) The CHB semantic similarity judging model was trained in the absence of video samples; and 2) The precision rate of this model is significantly affected by the HAR recognition model and the TAL model.

We meticulously examined the complete set of 14 CHB samples, including those that were incorrectly identified, which exhibited several inaccuracies in the extraction of basic human actions. From **Fig.11**, three noteworthy observations can be made:



(a)

(b)

(c)

(d)

Figure 11: Three typical experiment results.

(1)**Fig.11(a), (b)** illustrate two instances of correct recognition, although neither is flawless. Compared to the ground truth, one basic human action in each video sample was misidentified.

Nevertheless, the overall predicted results remain accurate. This suggests that the CHB semantic similarity assessment model we proposed exhibits strong fault tolerance. Since the word-vector is a crucial part of the input matrix, its vector nature means that only a small number of incorrectly extracted basic actions can minimally alter the direction and magnitude of the resultant vector derived from the complete set of action word-vectors and their durations. Such minor adjustments are unlikely to disrupt the classification thresholds between different CHBs.

(2)In a manner similar to the previous point, it is challenging, if not impossible, to obtain the SKF and EKF corresponding to each basic human action within a CHB sample. Biases are nearly present in all TAL results when compared to the ground truth. However, these discrepancies have minimal impact on the final results, thanks to the structure of the input matrix.

(3)**Fig 11(c)** shows a CHB sample that was incorrectly recognized. This case resembles Figures 11(a) and 11(b), as each of the three test results contains only one error derived from the basic human actions. However, Figure 11(c) yields an incorrect outcome, whereas the other two achieve correct results. This discrepancy arises because the SKF and EKF for the "make a phone call" action significantly deviated from their ground truth. Thus, even though the "make a phone call" action was accurately recognized, the final classification was incorrect.

5 Conclusions

In our work, we introduced a innovative method for recognizing Complex Human Behaviors (CHBs) using deep learning technology without the need for any training samples of video. Although the experimental accuracy achieved was only 71.43%, our methodology has been demonstrated to be feasible. A thorough review of the experimental outcomes confirms that our method possesses significant advantages in terms of universality and scalability, as it does not depend on CHB video training samples. Furthermore, the CHB model is capable of being efficiently retrained within a short period when new CHBs need to be identified.

A similarly significant discovery is that accuracy may be greatly improved through enhancing the effectiveness of the Human Action Recognition (HAR) model along with the accuracy of the Temporal Action Localization (TAL) model. As HAR and TAL technologies continue to advance, we anticipate that our proposed method's accuracy rate can progressively be increased.

However, a limitation of our proposed method is its strong dependency on the HAR and TAL models. Any biases in the outputs from these two models may lead to inaccuracies in the ultimate detection outcomes. As a result, our future efforts will primarily concentrate on consistently improving the effectiveness of the HAR model and the accuracy of the TAL model. Additionally, to improve the CHB model's capacity for similarity discrimination, we plan to incorporate more features of basic human actions into the input matrix.

Funding

This research was supported by the Science and Technology Research Projects of Jiangxi Provincial Department of Education under Grant GJJ2202613 and GJJ2402512, the National Natural Science Foundation of China under Grant 62366023, the Doctor Scientific Research Start-up Foundation of Jiangxi University of Chinese Medicine under Grant 2022BSZR007.

Author contributions

Qianglai Xie and Wei Lu: Extend and re-write the manuscript. Wei Yang: Software, Data Visualization. Keyun Xiong and Lei Zhang: Funding and data collecting. Leiyue Yao: Supervision, Reviewing and Editing.

Conflict of interest

No author associated with this paper has disclosed any potential or pertinent conflicts that may be perceived to have impending conflicts with this work.

References

[1] Z. Sun, Q. Ke, H. Rahmani, et al. Human Action Recognition from Various Data Modalities: A Review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 6: 1-20.

[2] Bakar A R. Advances in human action recognition: an updated survey. IET Image Processing. 2019, 13 (13): 2381 2394.

[3] Moustaka V, Vakali A, Anthopoulos L G. A systematic review for smart city data analytics. ACM Computing Surveys, 2018, 51(5): 1 41.

[4] Buzachis A, Celesti A, Galletta A, et al. A multi-agent autonomous intersection management (MA-AIM) system for smart cities leveraging edge-of-things and blockchain. Information Sciences, 2020, 522: 148 163.

[5] Kelly J W, Klesel B C, Cherep L A. Visual stabilization of balance in virtual reality using the HTC vive. ACM Transactions on Applied Perception, 2019, 16(2): 1 11.

[6] Rahimi Moghadam K, Banigan C, Ragan E D. Scene transitions and teleportation in virtual reality and the implications for spatial awareness and sickness. IEEE Transactions on Visualization and Computer Graphics, 2020, 26(6): 2273 2287.

[7] S. Majumder and N. Kehtarnavaz. Vision and inertial sensing fusion for human action recognition: A review. IEEE Sensors. 2021,21(3): 2454-2467.

[8] Ma X, Li Z, Zhang L. An Improved ResNet-50 for Garbage Image Classification. Tehnicki vjesnik - Technical Gazette, 2022, 29 (5):1552-1559.

[9] Macuzic S, Arsic B, Saveljic I, et al. Artificial Neural Network for Prediction of Seat-to-Head Frequency Response Function During Whole Body Vibrations in the Fore-and-Aft Direction.Tehnicki vjesnik - Technical Gazette, 2022, 29 (6):2001-2007.

[10] Ullah A, Muhammad K, Del Ser J, Baik SW, de Albuquerque VHC. Activity Recognition Using Temporal Optical Flow Convolutional Features and Multilayer LSTM. IEEE Transactions on Industrial Electronics 2019, 66: 9692-9702.

[11] Liu L, Shao L, Li X, Lu K. Learning Spatio-Temporal Representations for Action Recognition: A Genetic Programming Approach. IEEE Transactions on Cybernetics 2016, 46: 158-170.

[12] Ijjina EP, Chalavadi KM. Human action recognition in RGB-D videos using motion sequence information and deep learning. Pattern Recognition 2017, 72: 504-516.

[13] Zhang S, Chen E, Qi C, Liang C. Action Recognition Based on Sub-action Motion History Image and Static History Image. MATEC Web of Conferences 2016, 56: 2006.

[14] Abdelbaky A, Aly S. Human action recognition using short-time motion energy template images and PCANet features. Neural Computing and Applications 2020.

[15] Vishwakarma, Dinesh Kumar, Singh K. Human Activity Recognition Based on Spatial Distribution of Gradients at Sublevels of Average Energy Silhouette Images. IEEE Transactions on Cognitive & Developmental Systems 2017, 9(4):316-327.

[16] Arivazhagan S, Shebiah RN, Harini R, Swetha S. Human action recognition from RGB-D data using complete local binary pattern. Cognitive Systems Research 2019, 58: 94-104.

[17] Chen Y, Wang L, Li C, Hou Y, Li W. ConvNets-based action recognition from skeleton motion maps. Multimedia Tools and Applications 2019.

[18] Phyo CN, Zin TT, Tin P. Deep Learning for Recognizing Human Activities Using Motions of Skeletal Joints. IEEE Transactions on Consumer Electronics 2019, 65: 243-252.

[19] Ahmad T, Mao H, Lin L, Tang G. Action Recognition Using Attention-Joints Graph Convolutional Neural Networks. IEEE Access 2020, 8: 305-313.

[20] Caetano C, Bremond F, Schwartz WR. Skeleton Image Representation for 3D Action Recognition Based on Tree Structure and Reference Joints. 2019: 16-23.

[21] Liang X, Zhang H, Zhang Y, Huang J. JTCR: Joint Trajectory Character Recognition for human action recognition. 2019: 350-353.

[22] Wang X, Gao L, Wang P, Sun X, Liu X. Two-Stream 3-D convNet Fusion for Action Recognition in Videos With Arbitrary Size and Length. IEEE Transactions on Multimedia 2018, 20: 634-644.

[23] Wang Y, Sun J. Video Human Action Recognition Algorithm Based on Double Branch 3D-CNN. International Congress on Image and Signal Processing, BioMedical Engineering and Informatics. 2022.

[24] Li J, Liu X, Zhang M, et al. Spatio-temporal deformable 3D ConvNets with attention for action recognition. Pattern Recognition, 2020, 98: 107037-107045.

[25] Dai C, Liu X, Lai J. Human action recognition using two-stream attention based LSTM networks. Applied Soft Computing, 2020, 86: 105820-105827.

[26] Wang R, Luo H, Wang Q, Li Z, Zhao F, Huang J. A Spatial–Temporal Positioning Algorithm Using Residual Network and LSTM. IEEE Transactions on Instrumentation and Measurement, 2020, 69: 9251-9261.

[27] Li H, Shrestha A, Heidari H, Le Kernec J, Fioranelli F. Bi-LSTM Network for Multimodal Continuous Human Activity Recognition and Fall Detection. IEEE Sensors Journal 2020, 20: 1191-1201.

[28] Yadav N, Naik D. Generating Short Video Description using Deep-LSTM and Attention Mechanism. 2021: 1-6.

[29] Liu Z, Zhang H, Chen Z, et al. Disentangling and unifying graph convolutions for skeleton-based action recognition. IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp.143–152.

[30] Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition. AAAI conference on artificial intelligence, 2018.

[31] Shi L, Zhang Y, Cheng J, et al. Skeleton-based action recognition with directed graph neural networks. IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp.7912–7921.

[32] Chi H, Ha M, Chi S, et al. InfoGCN: Representation Learning for Human Skeleton-based Action Recognition. IEEE Conference on Computer Vision and Pattern Recognition, 2022.

[33] Xia H, Zhan Y. A Survey on Temporal Action Localization. IEEE Access 2020, 8: 70477-70487.

[34] Z. Shou, D. Wang, and S.-F. Chang, "Temporal action localization in untrimmed videos via multi-stage CNNs," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 1049–1058.

[35] Chao Y, Vijayanarasimhan S, Seybold B, et al. Rethinking the faster r-cnn architecture for temporal action localization. IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1130–1139.

[36] Lin T, Liu X, Li X, et al. Bmn: Boundary-matching network for temporal action proposal generation. IEEE International Conference on Computer Vision, 2019, pp. 3888–3897.

[37] Liu Y, Ma L, Zhang Y F, et al. Multi-granularity generator for temporal action proposal. IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp.3604–3613.

[38] Zhao P S, Xie L X, Chen J, Zhang Y, et al. Bottom-up temporal action localization with mutual regularization. IEEE International Conference on Computer Vision, 2020, pp. 539–555.

[39] Huang Y P, Dai Q, Lu Y T. Decoupling localization and classification in single shot temporal action detection. In ICME, 2019, pp. 1288–1293.

[40] Long F C, Yao T, Qiu Z F, et al. Gaussian temporal awareness networks for action localization. IEEE International Conference on Computer Vision, 2019, pp. 344–353.

[41] Wang H R, Yu B S, Xia K, et al. Skeleton edge motion networks for human action recognition. Neurocomputing, 2021:1-12

[42] Y. Du, Y. Fu, and L. Wang, Skeleton based action recognition with convolutional neural network, IEEE Asian Conference on Pattern Recognition, 2015: 579-583.

[43] Z. Yang, Y. Li, J. Yang, J. Luo, Action Recognition with Spatio-Temporal Visual Attention on Skeleton Image Sequences, IEEE Transactions on Circuits and Systems for Video Technology 2018. 29 (8):2405-2415.

[44] Yao L, Yang W, Huang W. A data augmentation method for human action recognition using dense joint motion images. Applied Soft Computing 2020, 97: 106713.

[45] Yao L Y, Yang W, Huang W, et al. Multi-scale feature learning and temporal probing strategy for one-stage temporal action localization. International Journal of Intelligent Systems, 2022, 6(1):1-10.

[46] T. Lin, X. Zhao, H. Su, C. Wang, and M. Yang, "BSN: Boundary sensitive network for temporal action proposal generation," in Proc. 15th Eur. Conf. Comput. Vis. (ECCV), Munich, Germany, Sep. 2018, pp. 3–21.

[47] T. Lin, X. Liu, X. Li, E. Ding, and S. Wen, "BMN: Boundary-matching network for temporal action proposal generation," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Seoul, South Korea, Oct. 2019, pp. 3888–3897.

[48] C. Lin et al., "Fast learning of temporal action proposal via dense boundary generator," in Proc. 34th AAAI Conf. Artif. Intell. (AAAI), New York, NY, USA, Feb. 2020, pp. 11499–11506.

[49] H. Su, W. Gan, W. Wu, Y. Qiao, and J. Yan, "BSN++: Complementary boundary regressor with scale-balanced relation modeling for temporal action proposal generation," in Proc. 35th AAAI Conf. Artif. Intell. (AAAI), Feb. 2021, pp. 2602–2610.

Cite this paper as:

Xie, Q.; Lu, W.; Yang, W.; Xiong, K.; Zhang, L.; Yao, L. (2025). Recognizing A Complex Human Behaviour via A Shallow Neural Network with Zero Video Training Sample, *International Journal of Computers Communications & Control*, 20(5), 6882, 2025.
https://doi.org/10.15837/ijccc.2025.5.6882