**CAMPUS ABNORMAL BEHAVIOUR RECOGNITION WITH TEMPORAL SEGMENT TRANSFORMERS**

**ABSTRACT**

The intelligent campus surveillance system is beneficial to improve safety in school. Abnormal behaviour recognition, a field of action recognition in computer vision, plays an essential role in intelligent surveillance systems. Computer vision has been actively applied to action recognition systems based on Convolutional Neural Networks (CNNs). However, capturing sufficient motion sequence features from videos remains a significant challenge in action recognition. This work explores the challenges of video-based abnormal behaviour recognition on campus. In addition, a novel framework is established on long-range temporal video structure modelling and a global sparse uniform sampling strategy that divides a video into three segments of identical durations and uniformly samples each snippet. The proposed method incorporates a consensus of three temporal segment transformers (TST) that globally connects patches and computes self attention with joint spatiotemporal factorization. The proposed model is developed on the newly created campus abnormal behaviour recognition (CABR50) dataset, which contains 50 human abnormal action classes with an average of over 700 clips per class. Experiments show that it is feasible to implement abnormal behavior recognition on campus and that the proposed method is competitive with other peer video recognition in terms of Top-1 and Top-5 recognition accuracy. The results suggest that TST-L+ can improve campus abnormal behavior recognition, corresponding to Top-1 and Top-5 accuracy results of 83.57% and 97.16%, respectively.

**INTRODUCTION**

Campus abnormal behavior recognition refers to using surveillance devices and artificial intelligence to identify unusual or potentially threatening behavior on campus. Video understanding is a core technology [1], [2], [3], [4] in many scenarios of surveillance systems. Over the years, unexpected actions, such as fighting, accidents, falling, and suicides, have occurred frequently in schools, causing general concern. Recognizing abnormal behavior can achieve real-time and efficient warning, positively affecting school safety management. Researchers focus on directly exploring abnormal behaviors instead of relying heavily on pre-processing to classify video behaviors [5], [6]. Researchers have focused on applications in specific scenarios on campus, such as classrooms [45] and laboratories [46], [47]. However, there is little research on campus abnormal behavior recognition. Essentially, abnormal behavior is a wide range of applications of video understanding. Motivated by video understanding, this study aims to provide an effective solution for recognizing video-based abnormal behavior on campus.

Deep learning has become increasingly popular for image recognition [7], [8], [9], [10], [11]. Depending on deep learning, video understanding has emerged in an endless stream to push action recognition to a climax. However, there are challenges in representing temporal video features, mainly focusing on three popular categories: (1) two-dimensional (2D) networks, (2) three-dimensional (3D) networks, and (3) transformers. In the first category, 2D network success represented by two-stream networks [12] pushed video understanding into the deep learning era. The following versions [13], [14], [15] related to two-stream networks emerged within a year, which have a similar network structure. One spatial network branch learns spatial information; the other is an optical flow network representing temporal information. Two-stream networks exhibit superior performance in learning spatial and temporal features separately. Because of the complex optical flow calculation and high storage requirements for pre-processing [12], previous studies are unsuitable for large-scale training and real-time deployment.

Meanwhile, for the second category: 3D networks, video understanding is a 3D tensor composed of two spatial dimensions and one temporal dimension to extract spatial and temporal features. However, optimizing the 3D model requires more work and relies heavily on diverse data than 2D networks [12], [16]. This situation changes when an inflated 3D model (I3D) [17] develops. The I3D operation can inflate Image Net’s pre-trained 2D model to the corresponding 3D model, accelerating optimization. Research related to 3D convolutional neural networks (CNNs) followed the emergence of I3D [18], [19], [20], [21], [22]. The 3D network with natural temporal properties [16] and inflated operation [17], [18], [19] have a competitive effect on video recognition. Consequently, it has long-dominated action recognition.

In the third category, transformers [23], [24], [25], [26] challenge the dominance of CNNs in deep learning and break the barriers of computer vision and natural language processing models. Because of its excellent capabilities in capturing distant information, especially for medium-range and long-range video modeling . Therefore, researchers are interested in applying transformers from the image field to video understanding [27], [28], [29]. However, if an element of self-attention consists of each image pixel, self-attention cannot be directly calculated in a transformer model with relatively high complexity. Hence, a fundamental problem to be solved is reducing the sequence length and designing a self-attention method. Times former [27] applied this sequence of frame-level patches with a size of 16 × 16 pixels instead of every pixel in an image and explored five structures of self-attention. This demonstrates that the divided space time attention method is faster to train than the 3D CNNs. Applying a transformer in video understanding on the kinetics 400 dataset achieved the best performance compared to CNNs [22] for the first time. However, the transformer has a common issue [23]: it is challenging to learn inductive bias owing to the lack of a large amount of pre-training data, like prior knowledge of the locality and translation equivariance in CNNs. Therefore, researchers have attempted to solve the inductive bias problem of pure transformers [28], [31], [32].

We aim to solve the problem of abnormal campus behavior recognition. Comparing different approaches with the results of the original paper above is challenging because a generic suite is needed to test these types of solutions using the new campus anomalous behavior dataset. Therefore, the proposed models are adequately compared with three relevant proposals: TSN [13], Slowfast [22], and Swin-B [31]. In addition, this work attempts to innovate abnormal behavior identification on campus. First, the backbone network consists of video shifted windows transformer [31], which effectively overcomes the inductive bias problem of the transformer: locality and translation equivariance. It also dramatically resolves the transformer sequence length issue and improves the global modeling ability of models by using a multi-scale shift window to calculate self-attention.

However, the problem with current models is their inability to model an entire video [13], [14]. Since they operate only on a single frame or a stack of frames within a short segment, they have limited access to the temporal context. The complex actions are illustrated in Fig. 1. Abnormal campus actions contain multiple segments with similar redundancies between the consecutive frames of one segment. Failure to use a long-range temporal structure for network training loses the ability to model the entire behavior. Motivated by this early work on video segmentation fusion TSN [13], we designed a campus abnormal behavior recognition framework called temporal segment transformer (TST) to exploit temporal action features and achieve video-level global modeling. Therefore, instead of working on a single frame or stacked frames, TST processes a sequence of snippets globally and is sparsely sampled from the entire video. Each snippet produces its initial class prediction, and a consensus function between the snippets is exported as the final prediction to enable global video dynamic modeling. It can remove redundant information and increase the difference between the behavior classes. Moreover, extensive experiments and discussions support a comparative study of these three methods. Overall, the main contributions of this study are summarized as follows:

• We propose a consensus of three temporal segment transformers (TST) based on the video Swin transformer for the new campus abnormal behavior recognition (CABR50) dataset. It enhances the ability to capture motion sequences and model long-range abnormal behavior on campus.

• We perform extensive comparative experiments with state-of-the-art methods for recognizing abnormal campus behaviors. The results show that it is feasible and can improve the accuracy of abnormal campus behavior recognition. In addition, we demonstrate the performance comparison of the TST and previous methods onthe UCF-101 dataset. It indicates that our proposed TST model has acceptable generalization performance.

• Our research provides essential technical support for the identification and early warning of abnormal behavior on campus, which plays an essential role in intelligent campus surveillance systems.

In the rest of the paper, Section II summarizes the research methods related to our work. Section III outlines how the TST is employed for campus abnormal behavior recognition and introduces related components. Section IV presents the experimental results and discussion. Section V concludes the work and presents perspectives.

**SYSTEM ANALSIS**

**EXISTING SYSTEM**

Xie et al. [45] used spatiotemporal representations to learn the posture estimation of college students to identify abnormal behavior. They analyzed the behavior of sleeping and using mobile phones in the classroom. Other researchers [46], [47] have explicitly looked at abnormal behavior in the laboratory. Rashmi et al. [46] apply YOLOv3 to locate and recognize student actions in still images from surveillance video in school laboratories. Unlike Rashmi’s image recognition-based analysis of students’ abnormal behavior in the laboratory, Banerjee et al. [47] used video. They propose a deep convolutional network architecture to detect and classify the behavioral patterns of students and teachers in computer-enabled laboratories. The above works can significantly demonstrate recognition of abnormal behavior in specific scenarios on campus. However, do not aim to reveal the feasibility of numerous abnormal behaviors in multiple scenarios on campus. Although there is limited research on campus aberrant behavior recognition, it is a form of video understanding. The following is a review of video understanding research relevant to our work.

Although self-attention is added as a submodule to CNNs to improve this temporal modeling video understanding, the ability of remote modeling can be made more robust by applying a pure transformer. These algorithms [27], [29], [30] are related to decomposing spatiotemporal self-attention using different factorized methods. They achieved better results than previous pure CNN and methods for adding self-attention units to videos. However, depending on the global attention modeling, these methods lead to a geometric increase in complexity.

Subsequently, MVTs [28] and the video Swin transformer [31] presented the idea of multi-scale hierarchical modeling by calculating the self-attention of multi-scale windows, which is much lower than the computational complexity of global self-attention. Moreover, it exceeds the previous decomposition space-time modeling methods in terms of accuracy and efficiency. The latest research [32] proposed a multi-view transformer consisting of multiple independent encoders to represent different dimensional input views, fusing information across views through horizontal connections. Although they achieve state-of-the-art performance, their method relies on many unpublic datasets and trained views. It may limit their generalization performance when applied to new videos or tasks beyond their original training scope. Therefore, a fairer comparison is required. In the case of employing ImageNet-21K as pre-training data to initialize network weights, the transformer method established on multi-scale hierarchical modeling [31] has more competitive advantages in video understanding.

**DISADVANTAGES**

* In the existing work, the system did not find Sensors and wireless data transmission for measuring food safety.
* This system is less performance due to lack of Real-time Prediction Campus Abnormal Behaviour.

**PROPOSED SYSTEM**

• We propose a consensus of three temporal segment transformers (TST) based on the video Swin transformer for the new campus abnormal behavior recognition (CABR50) dataset. It enhances the ability to capture motion sequences and model long-range abnormal behavior on campus.

• We perform extensive comparative experiments with state-of-the-art methods for recognizing abnormal campus behaviors. The results show that it is feasible and can improve the accuracy of abnormal campus behavior recognition. In addition, we demonstrate the performance comparison of the TST and previous methods on the UCF-101 dataset. It indicates that our proposed TSTmodel has acceptable generalization performance.

• Our research provides essential technical support for the identification and early warning of abnormal behavior on campus, which plays an essential role in intelligent campus surveillance systems.

**ADVANTAGES**

* The system is more effective since it involves Convolutional Neural Networks (CNNs) method.
* The system finds more ADVANTAGES OF THE system which designed a campus abnormal behavior recognition framework called temporal segment transformer (TST) to exploit temporal action features and achieve video-level global modeling.

**SYSTEMREQUIREMENTS**

**HARDWARE REQUIREMENTS**

➢ Processor - Pentium–IV

➢ RAM - 4 GB(min)

➢ Hard Disk - 20 GB

➢ KeyBoard - Standard Windows Keyboard

➢ Mouse - Two or ThreeButton Mouse

➢ Monitor - SVGA

**SOFTWARE REQUIREMENTS:**

**Operating system :** Windows 7 Ultimate.

**Coding Language :** Python.

**Front-End :** Python.

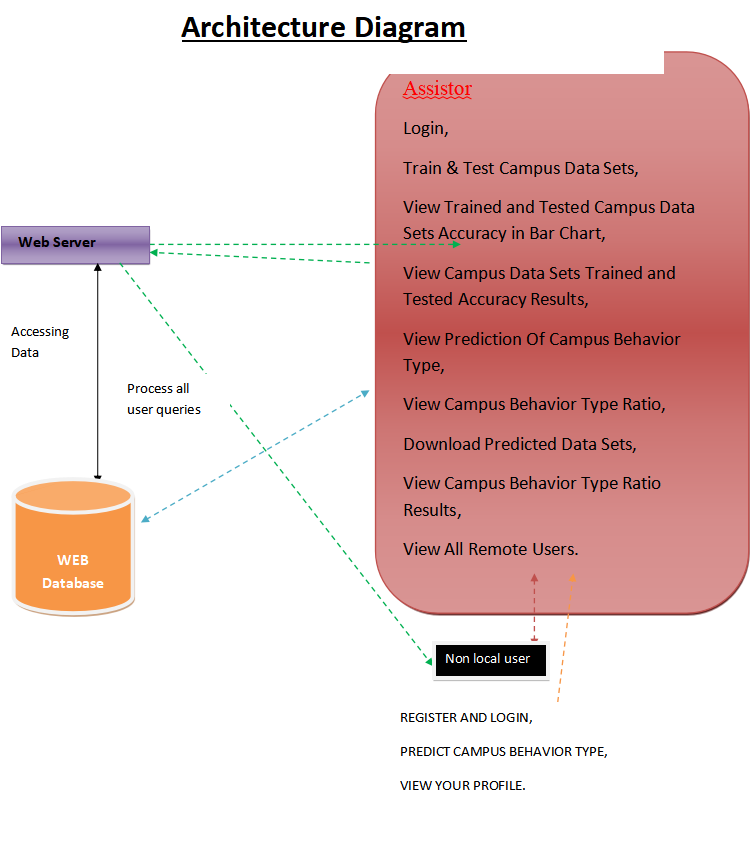
**Back-End :** Django-ORM

**Designing :** Html, css, javascript.

**Data Base :** MySQL

**MODULES**

* **ASSISTOR**
* **NONLOCAL USER**

**SYSTEM ARCHITECTURE **