CAMPUS ABNORMAL BEHAVIOR RECOGNITION WITH TEMPORAL SEGMENT TRANSFORMERS

[1]T. RAVI KIRAN KUMAR, [2]P. THRILOK,[3]G. JAGADEESH, [4]R. BALRAM GOUD, [5]D. VIJAY KUMAR,[6]R. GNANESHWAR REDDY

[1]ASSISTANT PROFESSOR, [2,3,4,5&6]UG STUDENTS

DEPARTMENT OF CSE, MNR COLLEGE OF ENGG. & TECHNOLOGY, MNR NAGAR, FASALWADIGUDA, SANGA REDDY-502294

ABSTRACT: The intelligent campus surveillance system is beneficial to improve safety in school. Abnormal behavior 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 behavior recognition on campus. In addition, a novel framework is established on long-range temporal video structure modeling 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 behavior 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.

1.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]. This situation changes when an inflated 3D model (I3D) 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. The 3D network with natural temporal properties and inflated operation have a competitive effect on video recognition. Consequently, it has long-dominated action recognition. In the third category, transformers 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. 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 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: 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. 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, and Swin-B. In addition, this work attempts to innovate abnormal behavior identification on campus. First, the backbone network consists of video shifted windows transformer, 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.

2.LITERATURE REVIEW

CAMPUS ABNORMAL BEHAVIOR RECOGNITION WITH TEMPORAL SEGMENT TRANSFORMERS HAIBIN LIU, JOON HUANG CHUAH, +2 AUTHORS X. WANG PUBLISHED IN IEEE ACCESS 2023 The intelligent campus surveillance system is beneficial to improve safety in school. Abnormal behavior 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 behavior recognition on campus. In addition, a novel framework is established on long-range temporal video structure modeling 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 behavior 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.

DIRECFORMER: A DIRECTED ATTENTION IN TRANSFORMER APPROACH TO ROBUST ACTION RECOGNITION THANH-DAT TRUONG, QUOC-HUY BUI, +4 AUTHORS KHOA LUU PUBLISHED IN COMPUTER VISION AND PATTERN… 19 MARCH 2022 Human action recognition has recently become one of the popular research topics in the computer vision community. Various 3D-CNN based methods have been presented to tackle both the spatial and temporal dimensions in the task of video action recognition with competitive results. However, these methods have suffered some fundamental limitations such as lack of robustness and generalization, e.g., how does the temporal ordering of video frames affect the recognition results? This work presents a novel end-to-end Transformer-based Directed Attention (Direc-Former) framework11The implementation of DirecFormer is available at https://github.com/uark-cviu/DirecFormer for robust action recognition. The method takes a simple but novel perspective of Transformer-based approach to understand the right order of sequence actions. Therefore, the contributions of this work are three-fold. Firstly, we introduce the problem of ordered temporal learning issues to the action recognition problem. Secondly, a new Directed Attention mechanism is introduced to understand and provide attentions to human actions in the right order. Thirdly, we introduce the conditional dependency in action sequence modeling that includes orders and classes. The proposed approach consistently achieves the state-of-the-art (SOTA) results compared with the recent action recognition methods [4, 18, 72, 74]. on three standard large-scale benchmarks, i.e. Jester, Kinetics-400 and Something-Something-V2.

DUALFORMER: LOCAL-GLOBAL STRATIFIED TRANSFORMER FOR EFFICIENT VIDEO RECOGNITION YUXUAN LIANG, PAN ZHOU, +1 AUTHOR SHUICHENG YAN PUBLISHED IN EUROPEAN CONFERENCE ON… 9 DECEMBER 2021 While transformers have shown great potential on video recognition with their strong capability of capturing long-range dependencies, they often suffer high computational costs induced by the self-attention to the huge number of 3D tokens. In this paper, we present a new transformer architecture termed DualFormer, which can efficiently perform space-time attention for video recognition. Concretely, DualFormer stratifies the full space-time attention into dual cascaded levels, i.e., to first learn fine-grained local interactions among nearby 3D tokens, and then to capture coarse-grained global dependencies between the query token and global pyramid contexts. Different from existing methods that apply space-time factorization or restrict attention computations within local windows for improving efficiency, our local-global stratification strategy can well capture both short- and long-range spatiotemporal dependencies, and meanwhile greatly reduces the number of keys and values in attention computation to boost efficiency. Experimental results verify the superiority of DualFormer on five video benchmarks against existing methods. In particular, DualFormer achieves 82.9%/85.2% top-1 accuracy on Kinetics-400/600 with ~1000G inference FLOPs which is at least 3.2x fewer than existing methods with similar performance.

VIDEO-CEPTION NETWORK: TOWARDS MULTI-SCALE EFFICIENT ASYMMETRIC SPATIAL-TEMPORAL INTERACTIONS YUAN TIAN, GUANGZHAO ZHAI, ZHIYONG GAO PUBLISHED IN ARXIV.ORG 22 JULY 2020 Previous video modeling methods leverage the cubic 3D convolution filters or its decomposed variants to exploit the motion cues for precise action recognition, which tend to be performed on the video features along the temporal and spatial axes symmetrically. This brings the hypothesis implicitly that the actions are recognized from the cubic voxel level and neglects the essential spatial-temporal shape diversity across different actions. In this paper, we propose a novel video representing method that fuses the features spatially and temporally in an asymmetric way to model action atomics spanning multi-scale spatial-temporal scales. To permit the feature fusion procedure efficiently and effectively, we also design the optimized feature interaction layer, which covers most feature fusion techniques as special case of it, e.g., channel shuffling and channel concatenating. We instantiate our method as a \textit{plug-and-play} block, termed Multi-Scale Efficient Asymmetric Spatial-Temporal Block. Our method can easily adapt the traditional 2D CNNs to the video understanding tasks such as action recognition. We verify our method on several most recent large-scale video datasets requiring strong temporal reasoning or appearance discriminating, e.g., Something-to-Something v1, Kinetics and Diving48, demonstrate the new state-of-the-art results without bells and whistles.

HIERARCHICAL TEMPORAL POOLING FOR EFFICIENT ONLINE ACTION RECOGNITION CAN ZHANG, YUEXIAN ZOU, G. CHEN PUBLISHED IN CONFERENCE ON MULTIMEDIA… 8 DECEMBER 2018 Action recognition in videos is a difficult and challenging task. Recent developed deep learning-based action recognition methods have achieved the state-of-the-art performance on several action recognition benchmarks. However, it is noted that these methods are inefficient since they are of large model size and require long runtime which restrict their practical applications. In this study, we focus on improving the accuracy and efficiency of action recognition following the two-stream ConvNets by investigating the effective video-level representations. Our motivation stems from the observation that redundant information widely exists in adjacent frames in the videos and humans do not recognize actions based on frame-level features. Therefore, to extract the effective video-level features, a Hierarchical Temporal Pooling (HTP) module is proposed and a two-stream action recognition network termed as HTP-Net (Two-stream) is developed, which is carefully designed to obtain effective video-level representations by hierarchically incorporating the temporal motion and spatial appearance features. It is worth noting that all two-stream action recognition methods using optical flow as one of the inputs are computationally inefficient since calculating optical flow is time-consuming. To improve the efficiency, in our study, we do not consider using optical flow but consider only raw RGB as input to our HTP-Net termed as HTP-Net (RGB) for a clear and concise presentation. Extensive experiments have been conducted on two benchmarks: UCF101 and HMDB51. Experimental results demonstrate that HTP-Net (Two-stream) achieves the state-of-the-art performance and HTP-Net (RGB) offers competitive action recognition accuracy but is approximately 1-2 orders of magnitude faster than other state-of-the-art single stream action recognition methods. Specifically, our HTP-Net (RGB) runs at 42 videos per second (vps) and 672 frames per second (fps) on an NVIDIA Titan X GPU, which enables real-time action recognition and is of great value in practical applications.

3.SYSTEMANALYSIS

3.1 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 OF EXISTING SYSTEM:

- 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

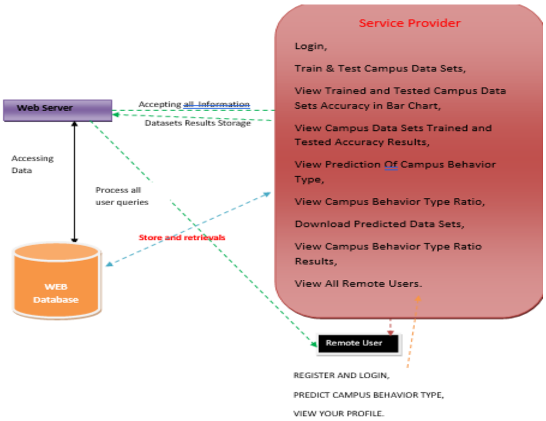
- We propose a consensus of three temporal segment transformers (TST) based on the video Swin transformer for the new campus abnormal behavior recognition (CABR50) dataset. It enhances the ability to capture motion sequences and model long-range abnormal behavior on campus.
- We perform extensive comparative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

ADVANTAGES OF PROPOSED SYSTEM:

- 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.

4.IMPLEMENTATION

4.1. SYSTEMARCHITECTURE


Service Provider
Login,
Train & Test Campus Data Sets,
View Trained and Tested Campus Data Sets Accuracy in Bar Chart,
View Campus Data Sets Trained and Tested Accuracy Results,
View Prediction Of Campus Behavior Type,
View Campus Behavior Type Ratio,
Download Predicted Data Sets,
View Campus Behavior Type Ratio Results,
View All Remote Users.
Web Server
Accepting all Information
Datasets Results Storage
Accessing Data
Process all user queries
Store and retrievals
WEB Database
Remote User
REGISTER AND LOGIN,
PREDICT CAMPUS BEHAVIOR TYPE,
VIEW YOUR PROFILE.

4.2. MODULES:

Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as Train & Test Campus Data Sets, View Trained and Tested Campus Data Sets Accuracy in Bar Chart, View Campus Data Sets Trained and Tested Accuracy Results, View Prediction Of Campus Behavior Type, View Campus Behavior Type Ratio, Download Predicted Data Sets, View Campus Behavior Type Ratio Results, View All Remote Users.

View and Authorize Users

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

Remote User

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN, PREDICT CAMPUS BEHAVIOR TYPE, VIEW YOUR PROFILE.

5.ALGORITHAMS

5.1. DECISIONTREE CLASSIFIERS

Decision tree classifiers are used successfully in many diverse areas. Their most important feature is the capability of capturing descriptive decision making knowledge from the supplied data. Decision tree can be generated from training sets. The procedure for such generation based on the set of objects (S), each belonging to one of the classes C1, C2, …, Ck is as follows:

Step 1. If all the objects in S belong to the same class, for example Ci, the decision tree for S consists of a leaf labeled with this class

Step 2. Otherwise, let T be some test with possible outcomes O1, O2,…, On. Each object in S has one outcome for T so the test partitions S into subsets S1, S2,… Sn where each object in Si has outcome Oi for T. T becomes the root of the decision tree and for each outcome Oi we build a subsidiary decision tree by invoking the same procedure recursively on the set Si.

5.2. GRADIENTBOOSTING

Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees.[1][2] When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees; it usually outperforms random forest.A gradient-boosted trees model is built in a stage-wise fashion as in other boosting methods, but it generalizes the other methods by allowing optimization of an arbitrary differentiable loss function.

5.3. K-NEAREST NEIGHBORS (KNN)

- Simple, but a very powerful classification algorithm
- Classifies based on a similarity measure
- Non-parametric
- Lazy learning
- Does not "learn" until the test example is given
- Whenever we have a new data to classify, we find its K-nearest neighbors from the training data

Example

- Training dataset consists of k-closest examples in feature space
- Feature space means, space with categorization variables (non-metric variables)
- Learning based on instances, and thus also works lazily because instance close to the input vector for test or prediction may take time to occur in the training dataset

6.RESULT

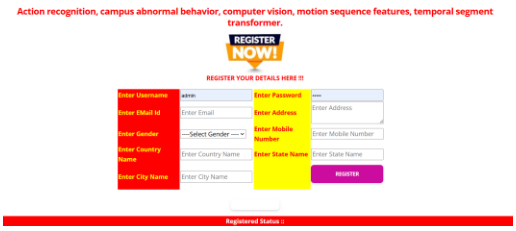
Campus Abnormal Behavior Recognition With Temporal Segment Transformers

Home| Remote User | Service Provider
MENTAL HEALTH

Action recognition, campus abnormal behavior, computer vision, motion sequence features, temporal segment transformer.

[illegible]

Home| Remote User | Service Provider

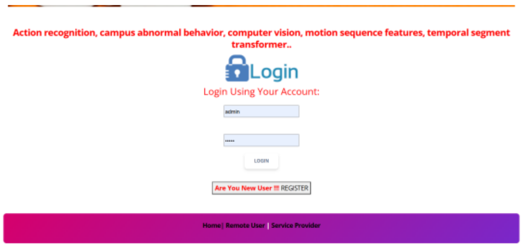


Action recognition, campus abnormal behavior, computer vision, motion sequence features, temporal segment transformer.

REGISTER YOUR DETAILS HERE !!!

Enter Username	admin	Enter Password	•••••
Enter EMail Id	Enter Email	Enter Address	Enter Address
Enter Gender	----Select Gender ----	Enter Mobile Number	Enter Mobile Number
Enter Country Name	Enter Country Name	Enter State Name	Enter State Name
Enter City Name	Enter City Name		REGISTER

Registered Status ::



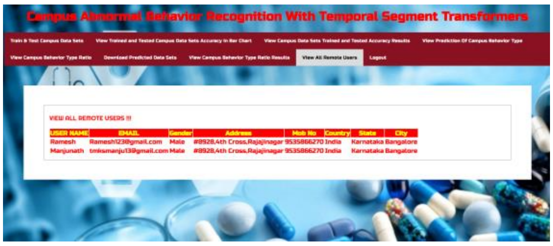
Action recognition, campus abnormal behavior, computer vision, motion sequence features, temporal segment transformer..

Login Using Your Account:

admin

•••••

LOGIN
Are You New User !!! REGISTER
Home| Remote User | Service Provider

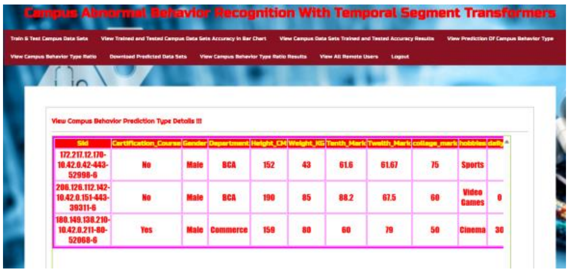


Campus Abnormal Behavior Recognition With Temporal Segment Transformers

Train & Test Campus Data Sets
View Trained and Tested Campus Data Sets Accuracy in Bar Chart
View Campus Data Sets Trained and Tested Accuracy Results
View Prediction Of Campus Behavior Type
View Campus Behavior Type Ratio
Download Predicted Data Sets
View Campus Behavior Type Ratio Results
View All Remote Users
Logout

VIEW ALL REMOTE USERS !!!

USER NAME	EMAIL	Gender	Address	Mob No	Country	State	City
Ramesh	Ramesh123@gmail.com	Male	#8928,4th Cross,Rajajinagar	9535866270	India	Karnataka	Bangalore
Manjunath	tmksmanju13@gmail.com	Male	#8928,4th Cross,Rajajinagar	9535866270	India	Karnataka	Bangalore

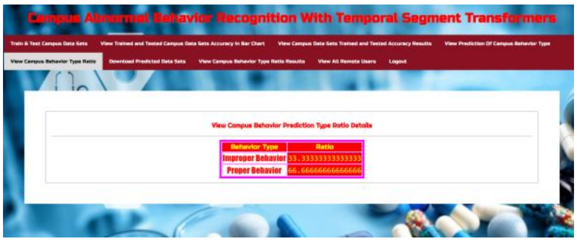


Campus Abnormal Behavior Recognition With Temporal Segment Transformers

Train & Test Campus Data Sets
View Trained and Tested Campus Data Sets Accuracy in Bar Chart
View Campus Data Sets Trained and Tested Accuracy Results
View Prediction Of Campus Behavior Type
View Campus Behavior Type Ratio
Download Predicted Data Sets
View Campus Behavior Type Ratio Results
View All Remote Users
Logout

View Campus Behavior Prediction Type Details !!!

Sid	Certification_Course	Gender	Department	Height_CM	Weight_KG	Tenth_Mark	Twelth_Mark	collage_mark	hobbies	daily
172.217.12.170-10.42.0.42-443-52998-6	No	Male	BCA	152	43	61.6	61.67	75	Sports	
206.126.112.142-10.42.0.151-443-39311-6	No	Male	BCA	190	85	88.2	67.5	60	Video Games	0
180.149.138.210-10.42.0.211-80-52068-6	Yes	Male	Commerce	159	80	60	79	50	Cinema	30

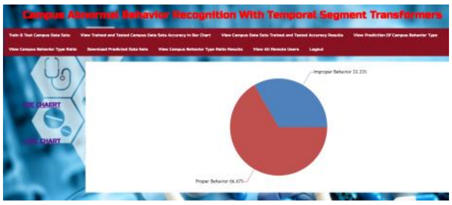


Campus Abnormal Behavior Recognition With Temporal Segment Transformers

Train & Test Campus Data Sets
View Trained and Tested Campus Data Sets Accuracy in Bar Chart
View Campus Data Sets Trained and Tested Accuracy Results
View Prediction Of Campus Behavior Type
View Campus Behavior Type Ratio
Download Predicted Data Sets
View Campus Behavior Type Ratio Results
View All Remote Users
Logout

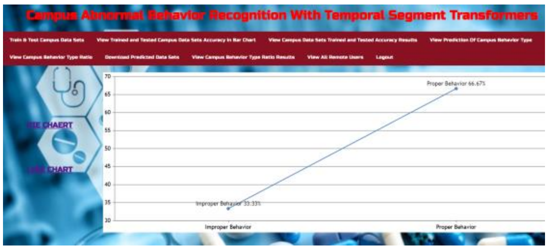
View Campus Behavior Prediction Type Ratio Details

Behavior Type	Ratio
Improper Behavior	33.333333333333336
Proper Behavior	66.66666666666666



Campus Abnormal Behavior Recognition With Temporal Segment Transformers

Train & Test Campus Data Sets
View Trained and Tested Campus Data Sets Accuracy in Bar Chart
View Campus Data Sets Trained and Tested Accuracy Results
View Prediction Of Campus Behavior Type
View Campus Behavior Type Ratio
Download Predicted Data Sets
View Campus Behavior Type Ratio Results
View All Remote Users
Logout
PIE CHAERT
LINE CHART
Improper Behavior 33.33%
Proper Behavior 66.67%



Campus Abnormal Behavior Recognition With Temporal Segment Transformers

Train & Test Campus Data Sets
View Trained and Tested Campus Data Sets Accuracy in Bar Chart
View Campus Data Sets Trained and Tested Accuracy Results
View Prediction Of Campus Behavior Type
View Campus Behavior Type Ratio
Download Predicted Data Sets
View Campus Behavior Type Ratio Results
View All Remote Users
Logout
PIE CHAERT
LINE CHART
70
65
60
55
50
45
40
35
30
Proper Behavior 66.67%
Improper Behavior 33.33%
Improper Behavior
Proper Behavior

CONCLUSION

Abnormal behavior recognition plays a significant role in intelligent campus surveillance systems. In this study, a CABR50 dataset was created, and a framework named temporal segment transformers was proposed to address the problem of identifying abnormal behavior on campus. Specifically, TST divides the input video into three segments of equal duration from the original video and obtains snippets uniformly sampled from its segment. These snippets are used as the inputs for the backbone network. Each snippet produces its initial prediction of the class, followed by a consensus function between the snippets exported as the final prediction, enabling dynamic global video modeling. Extensive experiments are carried out on the three split CABR50 datasets to verify the classification accuracy: TSN, Slow fast, Swin -B, and TST methods. In addition, the superiority of the proposed method in classifying abnormal behaviors on campus is verified in terms of the analysis result. To explore the model's generalization performance, we experimented with it on the UCF-101 dataset and achieved promising results. In summary, this work demonstrates the feasibility of using abnormal campus behavior recognition. In addition, our proposed TST can effectively model long-range behavior and achieve competitive results on CABR50. However, the model should perform better on the abnormal campus behavior categories of coughing, debating, and yelling that belong to hard data. In the future, we will try to combine multimodal approaches , such as extracting audio features from videos, to assist in classifying hard data for abnormal campus behavior. In addition, we can take an efficient approach to reduce the complexity of the model and overcome the problem of imbalance in the corresponding category data, such as GAN, which generates new training data for unbalanced categories.

REFERENCES

1. Q. Hao and L. Qin, ‘‘The design of intelligent transportation video processing system in big data environment,’’ IEEE Access, vol. 8, pp. 13769–13780, 2020.
2. K. Muhammad, J. Ahmad, Z. Lv, P. Bellavista, P. Yang, and S. W. Baik,‘‘Efficient deep CNN-based fire detection and localization in video surveillance applications,’’ IEEE Trans. Syst., Man, Cybern., Syst., vol. 49, no. 7, pp. 1419–1434, Jul. 2019.
3. L. Simoni, A. Scarton, C. Macchi, F. Gori, G. Pasquini, and S. Pogliaghi,‘‘Quantitative and qualitative running gait analysis through an innovative video-based approach,’’ Sensors, vol. 21, no. 9, p. 2977, Apr. 2021.
4. M. Shorfuzzaman, M. S. Hossain, and M. F. Alhamid, ‘‘Towards the sustainable development of smart cities through mass video surveillance: A response to the COVID-19 pandemic,’’ Sustain. Cities Soc., vol. 64, Jan. 2021, Art. no. 102582.
5. A. B. Mabrouk and E. Zagrouba, ‘‘Abnormal behavior recognition for intelligent video surveillance systems: A review,’’ Expert Syst. Appl., vol. 91, pp. 480–491, Jan. 2018.
6. A. B. Tanfous, H. Drira, and B. B. Amor, ‘‘Sparse coding of shape trajectories for facial expression and action recognition,’’ IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 10, pp. 2594–2607, Oct. 2020.
7. A. Krizhevsky, I. Sutskever, and G. E. Hinton, ‘‘ImageNet classification with deep convolutional neural networks,’’ in Proc. NIPS, Dec. 2012, pp. 1097–1105.
8. C. Szegedy,W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, ‘‘Going deeper with convolutions,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 1–9.
9. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, ‘‘Rethinking the inception architecture for computer vision,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 2818–2826.
10. K. He, X. Zhang, S. Ren, and J. Sun, ‘‘Deep residual learning for image recognition,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.
11. H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, M. Li, and A. Smola, ‘‘ResNeSt: Split-attention networks,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.Workshops (CVPRW), Jun. 2022, pp. 2735–2745.
12. K. Simonyan and A. Zisserman, ‘‘Two-stream convolutional networks for action recognition in videos,’’ in Proc. Adv. Neural Inf. Process. Syst., vol. 27, Dec. 2014, pp. 1–9.
13. L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. van Gool, ‘‘Temporal segment networks: Towards good practices for deep action recognition,’’ in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, Oct. 2016, pp. 20–36.
14. Z. Lan, Y. Zhu, A. G. Hauptmann, and S. Newsam, ‘‘Deep local video feature for action recognition,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jul. 2017, pp. 1219–1225.
15. A. Diba, V. Sharma, and L. Van Gool, ‘‘Deep temporal linear encoding networks,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1541–1550.