Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 5, No. 1, 2023

Research on Video Detection Method of Mudslide based on Inflated 3D Convolutional Neural Network

Zefeng Yu *

Chengdu University of Technology, Chengdu, China
*** Corresponding author Email:** yuzefeng2021@163.com

Abstract: Mudslide is a common natural disaster in mountainous areas, causing harm to roads and railroads and structures. To address this problem, this paper adopts the automatic video recognition approach, which utilizes widely installed video surveillance equipment to detect the changes of mudslides, so as to identify the mudslide diffuse flow disaster in the video monitoring area to achieve early warning. Firstly, the deep learning model is trained with weakly labeled mudslide video files, and the spatio-temporal feature learning method, i.e., inflated 3D convolutional network, is combined in the model, which results in a higher correctness rate of training and detection; secondly, the model is shown to have a recognition accuracy of 86% through the testing of the relevant datasets, which can be used as an effective complement to the traditional method of mudslide early warning.

Keywords: Debris Flow Transformation; Weak Labeling; Convolutional Networks.

1. Introduction

A mudslide is a natural disaster that occurs mainly in mountainous areas, on slopes or in areas following volcanic eruptions. It consists of a fluid flow of large quantities of rain, snowmelt or melted glacial water, as well as a mixture of sediment and rocks, which creates a destructive and powerful shock wave. Mudslides occur globally, but especially in mountainous areas, where they have a major impact on people, infrastructure and ecosystems [1].

Mudslides occur primarily when large amounts of rainfall, snowmelt, or glacial meltwater cause soil and rock on mountain slopes to become saturated and lose their support. When the amount of standing water exceeds the water-holding capacity of the soil, the soil becomes loose and joins with rocks and other debris to form fluid-like mudslides. In addition, natural disasters such as earthquakes, volcanic eruptions, or glacial lake outbursts can cause mudslides.

A mudslide is a devastating natural disaster and the damage caused is often enormous [2]. It can destroy homes, roads and bridges, leading to loss of life and property. At the same time, mudslides can block rivers, lakes and estuaries, creating mud and rock accumulations that can exacerbate the risk of flooding and inundation. In addition, mudslides can damage farmland, forests, and wildlife habitat, causing long-term impacts on ecosystems.

In order to mitigate the impacts of mudslides, many regions have adopted a range of preventive measures. These measures include establishing mudslide monitoring systems to monitor the occurrence and evolution of geologic hazards in real time and send out early warning messages; reducing the impact of mudslide impacts through engineering measures such as the construction of protective dikes, embankments, and berms; reinforcing slopes to reduce the formation of mudslides through the restoration of vegetation and ecological remediation; avoiding the construction of residential areas or important infrastructures in areas with high risk of mudslides; and strengthening the public's mudslide risk awareness and safety education to improve the ability to cope with disasters.

Currently, research on mudslides and disaster prevention is being carried out in many scientific research institutions and local governments. The research aspects involve the formation mechanism of mudslide, improvement of monitoring technology, and evaluation of the effectiveness of prevention and control measures. Computer technology also plays an important role in mudslide research, including mudslide simulation and prediction, mudslide monitoring and early warning systems, land planning and disaster risk assessment, and data analysis and early warning decision support. Through global cooperation and scientific and technological progress, it is expected that the capacity for early warning and prevention of mudslides will continue to improve, with the prospect of reducing the damage caused by the disaster.

2. Video Detection Principle

2.1. Anomaly Detection

In computer vision, anomaly detection is a key data analysis task and one of the most challenging and long-standing problems. In order to ensure that the model used in this paper can correctly identify abnormal behaviors, a large amount of data needs to be collected for training [3]. But collecting all the normal behavior data is almost impossible. At the same time, the boundary between normal and abnormal is blurred, and the determination of abnormality requires a reasonable grasp of the scale of normal. In order to solve the above problems, this paper adds some abnormal information for the model to make reference, i.e., not only the normal behavior is considered, but also the abnormal behavior is taken into account for the abnormality detection, and only the weakly labeled training data is used, so as to improve the model's understanding of the abnormal behavior.

And anomaly detection for video surveillance applications, in this paper is used for the monitoring of mudslide impact. In the testing process, patterns with large reconstruction errors are considered as anomalous behaviors. And in this paper, the action of mudslide is considered as the target of anomaly detection. Unlike the generic detection of still images, video detection usually utilizes the context between consecutive

frames of the video to localize the region of interest. In this paper, the model is trained under weak supervision, and thus only cares about the presence of anomalous events, not about which frames of the video the specific anomaly occurs within. Traditional action recognition methods cannot be used for anomaly detection in realistic surveillance videos. This is because the dataset contains long periods of untrimmed video, whereas anomalies tend to be random and mostly occur within a short period of time. As a result, the features extracted from these untrimmed training videos are not sufficiently discriminative for anomalous events. Therefore, training of abnormal and normal videos is indispensable. In this regard, in this paper, the videos in the dataset are all categorized into short videos and the different short videos are put into positive and negative example packages for subsequent operations.

2.2. I3D

In the field of image processing, all the images that are convolved are static images, so the use of two-dimensional convolutional neural network (2D CNN) is sufficient. In the field of video understanding, on the other hand, in order to retain the temporal information at the same time, it is necessary to learn the spatio-temporal features at the same time, and if a 2D CNN is used to process the video, then it will not be able to take into account the motion information encoded between multiple consecutive frames. For 2D convolutional networks, the input is an image and the output will only produce an image, and the input is a video and the output will still be an image, this is because 2D convolutional networks can only learn spatial features. In the previous use of 2D CNN to process video, the features are extracted from each key frame, and then an algorithm is used to combine the features of each key frame together. This operation creates a problem, that is, when using 2D CNN to process video is to treat each frame as a static picture, ignoring the time dimension of the motion information. 3D convolution and 3D pooling, on the other hand, can model temporal information, i.e., 3D convolutional network input video will output another video, preserving the input temporal information [4].

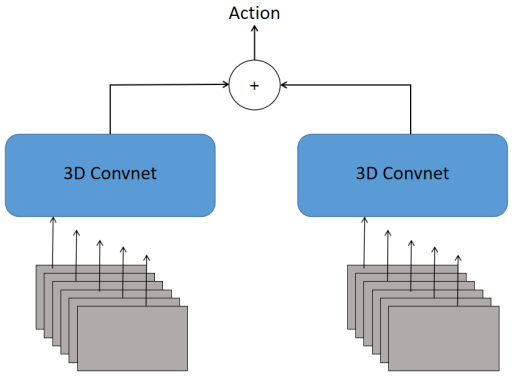

Fig 1. Schema of I3D

Meanwhile, in order to ensure that the network can extract features at multiple scales, so as to understand the video content more comprehensively, this paper adopts the Inception structure. Therefore, therefore, this paper adopts Inflated 3D Convolutional Neural Network (I3D) for the characteristics of mudslide recognition as well as the consideration of recognition success rate, as shown in Fig. 1. In addition, the overfitting is reduced by the Deconvolutional layer introduced by I3D to improve the generalization ability of the model, and the richness of the feature representation is enhanced by fusing different levels of features through feature splicing.

I3D is an innovative architecture in the field of deep learning, designed specifically for video data processing. Its core innovation is to extend the 2D convolutional weights, which have been subjected to an inflation operation, into 3D convolutional weights, thus enabling the network to capture both spatial and temporal dimensional information of video data.I3D is mainly used for tasks such as video classification and action recognition, and its construction and working process can be divided into several key steps.

First, I3D selects a convolutional neural network pre-trained on 2D image data as its infrastructure. This base network performs well in the domain of still images, but lacks the ability to process in the temporal dimension. Then, through a weight inflation operation, I3D expands these 2D convolutional weights into 3D convolutional weights suitable for the time dimension. This inflation operation transforms the 2D weights by adding an additional temporal dimension, allowing the network to perform convolutional operations in time. This allows I3D to better capture temporal information and dynamic changes in the video data. Next, the input is a video clip consisting of a series of consecutive video frames that are closely connected in time and represent video content over a period of time. These frames are preprocessed and normalized to serve as inputs to I3D [5].

In the core part of the network, i.e., the 3D convolutional layer, I3D uses an inflated 3D convolutional kernel to perform a convolutional operation on the sequence of input video frames. This allows the network to capture the motion and dynamics information of the video in the time dimension, giving it an advantage when dealing with spatio-temporal features such as actions and poses. Pooling and fully-connected layers are used to further compress the dimensionality of the feature maps and generate higher-level feature representations, which help the network better understand the abstract features of the video content.

Finally, I3D adds a classifier layer on top to map the network output to different categories or action labels. For the binary classification problem, this paper uses a Sigmoid activation function to obtain the probability that the representation belongs to a positive class, as shown in Figure 2. This allows I3D to provide accurate classification predictions for the input video clips.

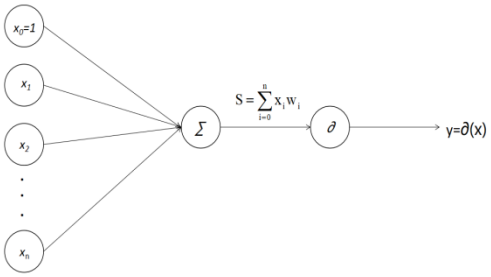

Fig 2. Sigmoid activation function example

The strength of I3D lies in its capability and flexibility. By extending 2D weights to 3D weights, I3D is able to process both spatial and temporal information of a video in a single network without additional processing steps. This allows it to better capture motion and dynamic patterns in videos, providing excellent performance for tasks such as video classification and action recognition. In the training phase, I3D can start with pre-trained weights from the image domain

and take advantage of migration learning to accelerate model convergence and improve performance.

2.3. Multi-example Learning

In terms of the ambiguity of the training data, there are three broad learning frameworks: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning has sample examples labeled; unsupervised learning does not. And multi-sample learning can be considered as a new kind of product that is different from the three major learning frameworks, and the mudslide detection studied in this paper precisely utilizes the method of multi-sample learning [6].

The usual deep learning training is a sample corresponds to a label, while in multiple examples learning, the concept of packet appears, one corresponds to a label, a positive packet needs at least one positive sample, a negative packet can only be all negative samples, and a packet, containing multiple samples. In this paper, the original dataset is preprocessed, the video is divided into several short frames, and according to whether there is anomaly collection into packets and labeling, respectively.

The purpose of multi-example learning is to build a multi-example classifier using multi-example packages with labels, and apply the classifier to the prediction of other unknown multi-example packages.

In this case, there are two types of prediction classes, one is package prediction and the other is example prediction. The biggest difference between the two prediction classes is the cost to the model after an example prediction error. For packet classification, if a needed feature is found in a packet, then it is classified as a positive example packet, and the features of the other examples in the packet are not concerned. Therefore, at this point, FP (judged as a positive sample, which is actually a negative sample) and FN (judged as a negative sample, which is actually a positive sample) have no effect on the accuracy of packet classification, but increase the example classification error rate. And when considering negative packages, an FP will allow a package to be misclassified (as long as there is a feature in the negative package that is needed for a positive example, the package will be classified as a positive package) [7].

In this paper, the short videos that have been segmented are assembled into packets and labeled, and then fed into the training model, as shown in Fig. 3. A multi-example classifier is built by learning multi-example packages with classification labels, and the classifier is applied to the prediction of unknown multi-example packages. Using multiple example packages for training can be more convenient and simpler.

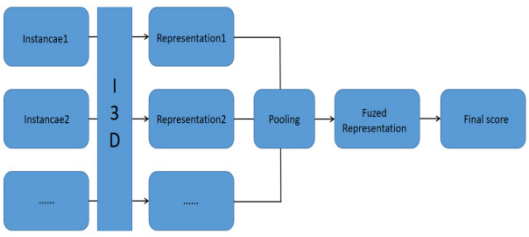

Fig 3. Multi-example learning

3. Convolutional Network-based Detection Methods

The cause of mudslide disaster is a solid-liquid two-phase fluid with a large number of solid materials such as mud, sand and stones, etc., formed due to a large amount of precipitation for a short period of time (heavy rainfall, snowmelt), which is in the state of viscous laminar flow or dilute turbulence, and it is a mixture of particulate flow of high concentration of solids and liquids [8]. Therefore, it is only necessary to look for anomalies in the surface in the dataset i.e., the presence of mudflow is clearly observed in the video.

This section explains the methodology for detection of debris flow in the video. After obtaining the monitored video, the image frames are extracted from the video file, which can be done in various ways depending on the requirements of the model. In this paper, the video processing library is used for this purpose. There is a library (OpenCV) in Python which is specialized for video processing. In this paper, this library is used to process the video file by reading the video frame by frame and setting the relevant parameters, after which each frame is saved as an image, which can be fed as an input to the deep learning model I3D in this paper for further processing and analysis.

3.1. Data Processing

When using a video processing library to process a video, it is important to set relevant parameters, such as the number and frequency of extracted video frame sequences.

Among them, the number of frames should be sufficient to capture the different stages of the action, but not too many to avoid excessive computational overhead. Therefore, in this paper, fifteen is taken as the number of frames to be extracted, and also to ensure the accuracy of the processing, this paper chooses to extract frames uniformly from the beginning, middle, and end parts of the video in order to obtain more comprehensive information.

Secondly, the frequency of extracted frames generally depends on the frame rate of the video and the time-dynamic information that wants to be captured. To characterize the data at the time of the mudslide disaster, this paper decides to use the key frame extraction method. An algorithm (optical flow method is used in this paper) is used to detect key frames in the video to capture significant changes.

3.2. Detection Principle

In this paper, we will train the model under weak supervision, i.e., in a video, we are only concerned with the presence or absence of anomalous events, without caring about the specific type of anomaly and the frames in which the anomaly occurs. Also, because the video captured by the camera has a certain degree of noise as well as light and shadow variations that can cause a certain amount of error in the results of the data, this paper uses the I3D training model to extract video features and the extracted features are used to calculate the anomaly score using the Sigmoid activation function, and based on the anomaly score, it is predicted whether or not an anomalous event occurs.

3.3. Model Building

Each training video is divided into the same number of clips to form positive and negative example packages for training. Then pick a best-scoring clip from the positive example packet for training the parameters of I3D, and similarly, choose a best-scoring clip from the negative example packet for training I3D.In this paper, we use the Cross-Entropy Loss function.

$$C = -(y \log(\hat{y}) + (1-y) \log(1-\hat{y}))$$

The cross-entropy loss function serves to measure the difference between the model's output probability distribution and the actual target distribution [9]. The smaller this difference is, the lower the loss value is, indicating the more accurate the model's prediction is. At the same time, the cross-entropy loss function is relatively simple for calculating the gradient of the model parameters, which allows optimization algorithms such as gradient descent to effectively update the parameters to reduce the loss. It helps the model to learn the correct category distribution and improve the classification accuracy.

3.4. Model Training

The videos in the dataset are divided into two parts according to 3:2 for training set and test set. After that the videos are divided equally without repetition into several short videos, each containing 15 frames of images, and such short videos are used as samples for training. The visual features are extracted using the fully connected layer of the I3D network. Before computing the features, each video frame was resized to 210 × 100 pixels and the frame rate was fixed to 30 fps. the I3D features for each 15-frame video clip were computed and then normalized. In order to obtain the features of a video clip, the average of the features of all 15-frame clips within that clip was taken. These features are later fed into the subsequent I3D neural network.

3.5. Results

The dataset as both normal and abnormal videos are labeled and sent to I3D for training gives very clear results as shown in Table 1. For video recognition, 10 videos were used from each event and they were divided into 3:2 ratio for training and testing. There is also another portion of the dataset trained using other models and a data comparison is done and the results are found to be very different. This is due to the fact that these are long untrimmed videos with low resolution and the anomalies appear very randomly. In addition, there are large variations due to changes in camera viewpoint and lighting as well as background noise [10]. Therefore, pre-processing the data would have made the experimental results more in line with expectations.

Table 1. Comparative results table

method	Methodology of this paper	C3D+CNN	C3D+MLP	TWO-STREAM
correct rate	86	83.0	76.3	64.0

4. Conclusion

Mudslide disasters can bring serious disaster impacts. In this paper, based on the video monitoring data, a combination of anomaly detection and I3D network is used to realize the detection of mudslide in the video. The occurrence of mudslide can be detected with high accuracy, thus realizing the disaster warning of mudslide and diffuse flow area range, providing a new idea for mudslide disaster monitoring and warning, and enriching the mudslide disaster prevention and mitigation system.

References

[1] Lohumi K, Roy S. Automatic detection of flood severity level from flood videos using deep learning models[C]//2018 5th International Conference on Information and Communication Technologies for Disaster Management (ICT-DM). IEEE, 2018: 1-7.

[2] Lopez-Fuentes L, van de Weijer J, Bolanos M, et al. Multi-modal Deep Learning Approach for Flood Detection[J]. MediaEval, 2017, 17: 13-15.

[3] Pang G, Shen C, Cao L, et al. Deep learning for anomaly detection: A review[J]. ACM computing surveys (CSUR), 2021, 54(2): 1-38.

[4] Soltani Nejad S. Weakly-Supervised Anomaly Detection in Surveillance Videos Based on Two-Stream I3D Convolution Network[J]. 2023.

[5] Munukutla P S, Jain S. One Shot Learning for Video Object Segmentation using Fully Convolutional I3D Network[J].

[6] Kexuan W, Yifei X, Xinyu X. Research on Machine Learning Based on Multi-label Algorithm[C]//2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA). IEEE, 2020: 824-829.

[7] Huaihui C, Shengbo Z. Multi Instance Deep Learning Target Tracking[J]. The Frontiers of Society, Science and Technology, 2020, 2(9).

[8] Chaganti P C V, Vasireddy K, Reddy E R, et al. Predicting Landslides and Floods with Deep Learning[C]//2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC). IEEE, 2023: 1259-1265.

[9] Li L, Doroslovački M, Loew M H. Approximating the gradient of cross-entropy loss function[J]. IEEE access, 2020, 8: 111626-111635.

[10] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3d convolutional networks[C]//Proceedings of the IEEE international conference on computer vision. 2015: 4489-4497.