Video Classification based on the Behaviors of Children in Pre-school through Surveillance Cameras

Tran Gia The Nguyen[1,2], Pham Phuc Tinh Do[1,2], Dinh Duy Ngoc Cao[1,2], Huu Minh Tam Nguyen[1,2], Huynh Truong Ngo[1,2], and Trong-Hop Do[1,3]

[1] University of Information Technology, VNU-HCM, Vietnam
[2] {20521940,20522020,20521661,20521871,20522085}@gm.uit.edu.vn
[3] hopdt@uit.edu.vn

Abstract. In preschool, children are active and curious about the world around them. The tendency to engage in unusual or dangerous behaviors can pose a significant risk to their safety. Constantly monitoring surveillance camera footage and analyzing it to determine if any abnormal behavior is occurring requires considerable attention and effort from human observers. Therefore, applying technology to moni- tor the abnormal behavior of children is crucial in preschool education settings. With the development of technology, the problem of classifying video through surveillance cameras can be solved. However, there is still a shortage of datasets for this task with video data in preschool environments due to difficulties in collecting data and potential violations of children's privacy and safety. Therefore, in this paper, we propose Behaviors of Children in Preschool (BCiPS), a new dataset for the above problem. BCiPS consists of 4268 videos with lengths ranging from 3-6 seconds. We evaluate some machine learning and deep learning models on BCiPS. The CNN-LSTM model achieves the highest performance, over 75%, with four performance metrics: accuracy, recall, precision, and $F1_{score}$. Additionally, we will analyze cases where the models fail to identify abnormal behavior to determine the reasons for these fails- ures, identify weaknesses in the dataset, and improve the accuracy of the dataset to create a reliable tool for building an effective warning system in real-life situations.

Keywords: Video Classification, BCiPS, Pre-school, Deep Learning, Surveillance Cameras

1 INTRODUCTION

The purpose of action classification in videos is to determine what is happening in the video. Human Activities can be classified into various categories, including Human-Human interaction and Human-Object interaction. This classification is based on human actions specified by their gestures, poses, etc... Human action recognition is challenging due to variations in motion, illumination, partial occlusion of humans, viewpoint, and anthropometry of people

involved in the various interactions. Additionally, the issue of a person's style when performing a gesture, not only in timing but also in how to perform the gesture.

In recent years, advancements in neural network technology and deep learning have become effective methods for many tasks, including video action classification. Along with the emergence of pre-trained models, the task of classifying action has also significantly improved. However, a few datasets exist in the school domain, especially in preschool. Due to the problem of personal privacy, child protection is a sensitive issue, affecting many aspects of the lives of those recorded by the camera.

Understanding such a situation, we have created the BCiPS dataset, including videos on the preschool domain. Our dataset includes videos capturing the activities of children in preschool. BCiPS is a valuable resource for researchers, enabling them to conduct analyses and develop models to identify children's abnormal behaviors. Regarding legal matters, we have contacted relevant competent parties and obtained permission to provide videos from the preschool surveillance camera.

Therefore, in this study, we performed the task of video classification based on the behaviors of children in preschool through surveillance cameras. The input of the task is a video, and the output is the classification of normal or abnormal human actions. After completing the experiments and research, we achieved two results:

- Defined the problem and provided guidelines for creating the BCiPS dataset, which consists of 4,268 videos ranging from 3 to 6 seconds in length. BCiPS is one of the first datasets containing videos extracted from surveillance cameras in the preschool domain.
- Experimented with deep learning models for the video classification task on the BCiPS dataset. Furthermore, we evaluated the performance of the models. The best-performing model was CNN+LSTM, achieving an accuracy of 75.88%. Additionally, we analyzed error cases to identify challenging scenarios that could help future research avoid similar errors and improve the performance of the models for real-world applications.

The remaining parts of this paper are as follows. Section 2 is the section that introduces previous datasets for classifying human actions in videos. Section 3 presents the BCiPS dataset. Section 4 shows the baseline models used in the study. Section 5 describes the performances of models. Section 6 presents the achieved results and future works.

2 RELATED WORKS

Many datasets have been published for human action recognition in videos. These are large-scale datasets with many labels and various topics in various fields. Some examples of these datasets include HMDB-51: This dataset spans 51 action classes and contains 6,766 clips extracted from 3,312 videos. The

UCF101 [12] dataset contains 101 action classes grouped into five types of actions and a total of 13,320 clips extracted from 2,500 videos. Kinetics Human Action Video Dataset [5] has 400 action classes and 306,245 clips extracted from 306,245 videos. NTU RGB+D [10]: This human action recognition dataset in a school domain. It includes 60 action classes, divided into three groups: 40 daily actions, nine health-related actions, and 11 mutual actions. Kinetics-700 [2]: This dataset contains over 650,000 videos and 700 action classes, with an average length of 10 seconds per video. The data domain of the dataset includes sports actions, movies, online videos, etc. The dataset was collected from YouTube, sports, and movie websites. The Something-Something dataset [4] contains over 1000,000 videos with 174 action classes.

3 THE BCiPS DATASET

3.1 Dataset Creation

Phase 1. Data Collection and Pre-processing:

Data collection: We describe creating the BCiPS dataset for human action classification in videos task. The dataset was created by collecting video footage from two surveillance cameras in two preschool classrooms. The cameras recorded the daily activities of students during various times of the day, including studying, entertaining, and napping. Careful consideration was given to ensuring the privacy and safety of the children.

Data pre-processing: Video data were collected in .dav format and then converted to .mp4 format using the web tool 123APPS [1]. This format is widely used and easy to work with various tools. As the study focuses on human actions, we removed any periods when no students were in the classroom from the videos. Subsequently, we used the FFmpeg library provided by Python to split the videos into 3-6 seconds segments.

Phase 2. Guidelines and Agreement:

Guidelines: The purpose of the task is to classify whether the input video contains normal or abnormal actions. Therefore, the output will be one of two labels: "Normal" or "Abnormal". We define the labels as follows:

- **Abnormal:** Videos contain unusual and dangerous actions for children, such as fighting (children physically impact each other or play excessively, causing harm to other children), falling (falling on the ground suddenly), chasing (more than two students run fast, they run without a certain trajectory, make noise and may have an accident), carrying heavy or oversized objects (tables, chairs, beds, or similarly large-sized items can cause injury to children if they trip or if heavy objects fall on them), and other abnormal activities (other less common cases can

[1] https://123apps.com/

still pose dangers to children). These actions can be dangerous, leading to injury and accidents or affecting the children's mental health.
- **Normal:** Videos contain actions that are not dangerous in the observed environment. For example, walking, talking, eating, playing, studying, ..., performed normally and regularly in a school or preschool domain.

For videos that children are not sleeping, we propose watching the actions in the video and evaluating whether they fall within the range of normal actions. If the activities in the video do not belong to the list of abnormal actions and do not harm children, we will label the video as "Normal".

Fig. 1: Three frames of three videos labeled as 'Abnormal'.

Regarding the frame captured from CAM15, it shows a child carrying a heavy object (a chair) while another child is standing on the chair. This could lead to a fall and harm the children. As for the middle frame (left panel of CAM16), two children are chasing each other, which could also lead to falling and cause danger. In the frame on the right panel of CAM16, a girl is pulling a shirt and physically impacting someone else in the frame. Therefore, three videos are labeled as "Abnormal".

Annotators Agreement

To ensure the quality and objectivity of the dataset, we hired five annotators to label the data, and these annotators were provided with labeling guidelines. Each annotator independently labeled 150 identical videos to assess their labeling consistency and determine the official labels. We measured the inter-annotator agreement among the annotators using Cohen's Kappa index and the average annotation agreement [3]. According to the ranking table for annotation agreement in classification data [8], Table 1 shows that the average agreement among pairs is 0.68, which is considered good as it falls within the range of 0.6 and 0.81.

In addition to calculating the annotation agreement among each pair using Cohen's Kappa index, we also calculated the agreement among all five annotators using Krippendorff's alpha [7]. The agreement was 67.55%, which is higher than the acceptable value (66.7%) for Krippendorff's alpha. Therefore, after labeling the data according to the guidelines in section 3.1, our agreement was higher than the minimum acceptable result for Krippendorff's alpha and achieved good agreement according to Cohen's Kappa index. As a result, we can officially label the collected data.

Table 1: The agreement among annotators as measured by Cohen's Kappa.

	A1	A2	A3	A4	A5
A1	1	0.70	0.64	0.71	0.73
A2	-	1	0.46	0.56	0.56
A3	-	-	1	0.81	0.73
A4	-	-	-	1	0.86
A5	-	-	-	-	1
Average	0.68				

Phase 3. Labeling and data splitting: The inter-annotator agreement results for both Cohen's Kappa and Krippendorff measures were good, meeting the set requirements. Therefore, we proceeded to label the videos in our dataset. It is inevitable to encounter challenging and ambiguous cases during the labeling process. To address these issues, the annotators will have meetings to discuss and agree on how to label these problematic and ambiguous cases and update the guidelines to make them more complete. Following this process, we obtained a total of 4268 labeled videos. Finally, we split 4268 videos into three sets: training set, validation set (development set), and test set.

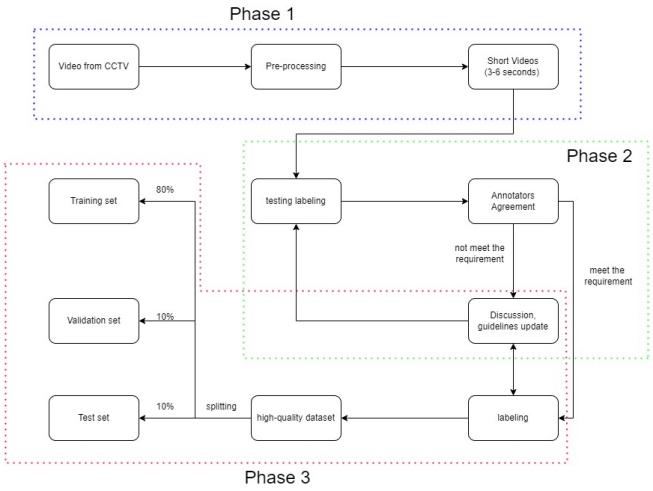

Fig. 2: Dataset creation processing

3.2 Dataset Analysis

After collecting, preprocessing, and labeling data, BCiPS include two labels, 4,268 videos with a total size of 19.9GB, and a total length of 21,353 seconds.

We divided the dataset into three sets: training, development, and testing, with a ratio of 8:1:1.

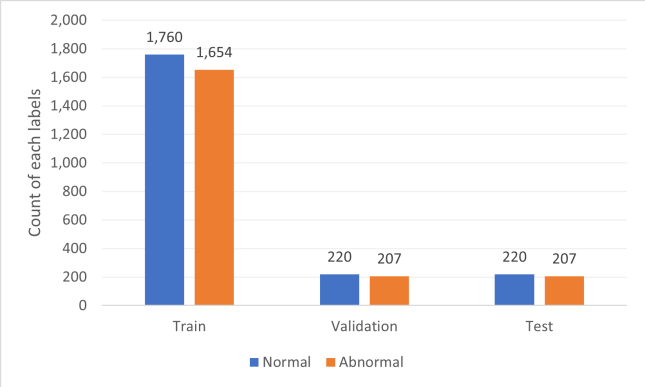

Fig. 3: Distribution of labels in the training, validation, and testing sets.

Figure 3 shows that the number of Normal and Abnormal labels is relatively balanced. In three sets, the ratio of 2 labels is relatively equal. The train set is much larger than the test and validation set because we want the model to learn more cases, covering the data. For the test set and the validation set with a ratio of 1:1, this helps evaluate the most objective and close to reality.

4 BASELINE MODELS

4.1 CNN+LSTM (ConvLSTM2D)

The Convolutional LSTM (ConvLSTM) Network was first introduced in the work of [11]. In a fully connected LSTM network, flattening the image into a 1D space does not retain any spatial information, hence the need for CNN to extract spatial features and transform them into a 1D vector space. Therefore, the ConvLSTM network was proposed for video classification tasks [9], using 2D structures as inputs. It can directly work with a sequence of images and perform convolutional operations on the input images to extract spatial features, while LSTM layers can extract temporal dynamics between frames. Therefore, the ConvLSTM network can capture spatial and temporal signals, which fully connected LSTM cannot achieve.

4.2 CNN+SVM

CNN is used to extract features from video data. It is typically used to learn 2D image features from each frame. The output results of CNN for each frame generate a feature vector, which is subsequently utilized to train the SVM model.

4.3 CNN+Random Forest

CNN is used to extract features from video data. It is typically used to learn 2D image features from each frame. The output results of CNN for each frame generate a feature vector. It is then used to train the Random forest model.

4.4 TimeSformer

As TimeSformer (Time-Space Transformer) [1] is one of the most advanced methods for video classification that has recently emerged, we have decided to use this architecture in our benchmark. The TimeSformer model is designed for videos and is pre-trained on the ImageNet dataset and fine-tuned on our dataset. TimeSformer does not contain convolutional layers. Instead, it employs a self-attention mechanism. TimeSformer adapts transformer architecture for computer vision and video processing tasks.

4.5 MoViNets

MoViNets [6] is a video classification model used for online video streaming or real-time inference in tasks such as action recognition. The classifier of MoViNets is based on efficient and simple 2D frame-level processing to run over the entire video or stream each frame one by one. As it cannot account for temporal context, it has limited accuracy and may give inconsistent output results from one frame to another. A simple 3D CNN that uses a two-dimensional temporal context can increase accuracy and temporal consistency.

4.6 (2+1)D Resnet-18

The following 3D convolutional neural network model is based on the work published by D. Tran et al.. The (2 + 1)D convolution allows for the separation of spatial and temporal dimensions, creating two separate steps. One advantage of this method is that it helps save parameters by analyzing combinations of spatial and temporal dimensions

4.7 EfficientNetB0

EfficientNet [13] was introduced by Tan and Le, who studied the scaling of models and determined that carefully balancing a network's depth, width, and resolution can lead to better performance. They proposed a novel scaling method that evenly scales all dimensions of a network, including depth, width,

and resolution. They used a neural architecture search tool to design a new base network and extended it to obtain a family of deep learning models. There are 8 variants of EfficientNet (B0 through B7), with EfficientNetB0 having 5.3 million parameters. Figure 5 illustrates the architecture of the EfficientNet network.

5 RESULTS

5.1 Results of Baseline Models

Table 2: Models Performance

	Accuracy	Recall	Precision	F1
CNN + LSTM	**75.88**	**75.63**	**75.43**	**75.53**
Timesformer	74.71	74.80	74.82	74.81
CNN + RandomForest	74.00	73.97	73.98	73.98
EfficientNetB0	73.30	73.48	73.70	73,59
Movinet	70.02	69.61	71.30	70.44
CNN + SVM	65.57	65.61	65.59	65.60
(2+1)D Resnet-18	61.12	60.42	63.37	61,86

Table 2 shows that the CNN-LSTM achieves the highest results in all four metrics: accuracy, recall, precision, and F1-score with 75.88%, 75.63%, 75.43%, 75.53% respectively. The results of the CNN-LSTM model are significantly different from the other models. Besides, the (2+1)D Resnet-18 model has the lowest performance with only 61.12%, 60.42%, 63.37%, and 61.86%, much lower than the other models. It can be seen that the pre-trained models resulted in lower performance with our dataset than those combining traditional models.

5.2 Error analysis

According to Figure 4, the precision of the "Normal" class is 73.31%, while the precision of the "Abnormal" class is 79.55%. This suggests that the model tends to classify videos as abnormal rather than ignoring abnormal videos. The difference in precision between the two classes indicates that the model is moving in the right direction for the development of the dataset, but further improvement is still required. The error rate and the difference in the precision of the two classes suggest that the model is still prone to misclassifications and needs to be fine-tuned for better accuracy.

Figure 5 shows that the video contains many ambiguous actions (e.g., "Running"), which caused the model to predict it as abnormal. The main reason is that some abnormal behaviors have not been fully defined. Additionally, the video also includes some ambiguous actions, and the dataset needs to be more diverse.

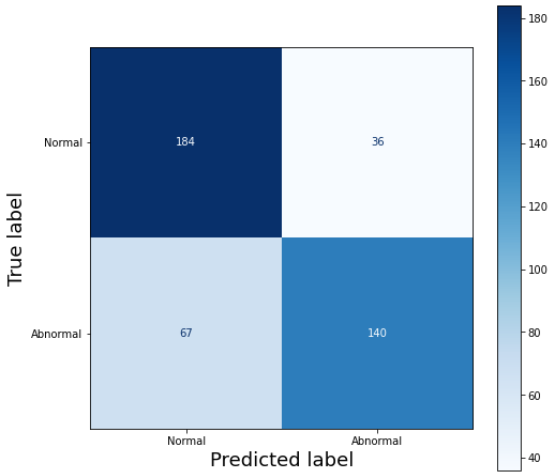

Fig. 4: Confusion Matrix

Fig. 5: Predicted label and actual label of a sample video from the test set.

6 Conclusion-Future Works

In this paper, we have created a dataset for classifying the behaviors of children in preschool named BCiPS. There are two labels, "Normal" and "Abnormal". After experimenting with the dataset on several video classification models, including basic, combined, and pre-trained models, we achieved the highest accuracy of 75.88% with the CNN + LSTM model. This demonstrates the potential of the dataset we have constructed for video classification in a preschool domain. However, specific errors still need to be addressed, which hinder the optimal performance of the models. In order to achieve real-world application, we plan to improve the dataset by collecting more data, re-labeling the data more effectively, and establishing stricter guidelines for data labeling.

Observing the existing limitations in the dataset, we intend to take some improvement steps towards applying the dataset in practical applications. Specif-

ically, we will collect additional data, re-label the dataset more efficiently, and establish stricter labeling guidelines. Finally, we will explore suitable models for this dataset.

References

1. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: ICML, vol. 2, p. 4 (2021)
2. Carreira, J., Noland, E., Hillier, C., Zisserman, A.: A short note on the kinetics-700 human action dataset. arXiv preprint arXiv:1907.06987 (2019)
3. Cohen, J.: A coefficient of agreement for nominal scales. Educational and psychological measurement **20**(1), 37–46 (1960)
4. Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., Hoppe, F., Thurau, C., Bax, I., Memisevic, R.: The "something something" video database for learning and evaluating visual common sense. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017)
5. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
6. Kondratyuk, D., Yuan, L., Li, Y., Zhang, L., Tan, M., Brown, M., Gong, B.: Movinets: Mobile video networks for efficient video recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16,020–16,030 (2021)
7. Krippendorff, K.: Content analysis (2004)
8. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. biometrics pp. 159–174 (1977)
9. Luo, W., Liu, W., Gao, S.: Remembering history with convolutional lstm for anomaly detection. In: 2017 IEEE International Conference on Multimedia and Expo (ICME), pp. 439–444. IEEE (2017)
10. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1010–1019 (2016)
11. Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. Advances in neural information processing systems **28** (2015)
12. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
13. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning, pp. 6105–6114. PMLR (2019)