UDC 004.89

DOI: 10.15587/1729-4061.2024.296500

DETERMINING THE PSYCHO-EMOTIONAL STATE OF THE OBSERVED BASED ON THE ANALYSIS OF VIDEO OBSERVATIONS

Yedilkhan Amirgaliyev
Doctor of Technical Sciences, Professor, Chief Researcher,
Head of Laboratory
Laboratory of Artificial Intelligence and Robotics
Institute of Information and Computational Technologies
Shevchenko str., 28, Almaty, Republic of Kazakhstan, 050010

Yurii Krak
Doctor of Physical and Mathematical Sciences, Professor,
Head of Department
Department of Theoretical Cybernetics
Taras Shevchenko National University of Kyiv
Volodymyrska str., 60, Kyiv, Ukraine, 01033

Indira Bukenova
Corresponding author
Master of Technical Sciences, Lecturer*
E-mail: indira.bukenova11@gmail.com

Bayan Kazangapova
Associate Professor

Gani Bukenov
Master of Mathematics, Lecturer*
Department of Technology**
*Department of Information Systems**
**Almaty Technological University
Tole-Bi str., 100, Almaty, Republic of Kazakhstan, 050012

This paper develops a system for determining the psycho-emotional state of the observed people based on the analysis of video surveillance with the application of artificial intelligence technology using hardware and software tools such as PoseNet, PyTorch, SQLite, FastAPI and Flask. In many areas of human endeavor, there is an urgent need for a surveillance system that can reliably function and detect suspicious activities. To solve this problem, this paper proposes a novel framework for a real-time surveillance system that automatically detects abnormal human activities.

The system has been tested and validated in real environments. The results of testing artificial intelligence program models showed the best results (f1 score with values of 0.98–0.99). The weighted average value of the f1-score metric was 0.96, which is quite a high value. The use of PoseNet implemented with PyTorch allowed to accurately determine the pose of the person in the video and extract information about the position of different body parts. The peculiarity of this work lies in the development of artificial intelligence models for automatic detection of possible physical aggression in videos, in the methods of forming an optimal set of features for the development of AI models that identify the aggressor and the victim of bullying.

The developed system has the potential to be a useful tool in various fields such as psychology, medicine, security and others where it is important to analyze the emotional state of people based on their physical manifestations. The obtained applied results can be used in educational institutions and in spheres where video analysis is necessary

Keywords: computer vision, physical aggression, emotional reaction, bullying, model training, neural network

Received date 31.11.2023
Accepted date 13.02.2024
Published date 28.02.2024

***How to Cite**: Amirgaliyev, Y., Krak, Y., Bukenova, I., Kazangapova, B., Bukenov, G. (2024). Determining the psycho-emotional state of the observed based on the analysis of video observations. Eastern-European Journal of Enterprise Technologies, 1 (2 (127)), 45–53. doi: https://doi.org/10.15587/1729-4061.2024.296500*

1. Introduction

The development of automated systems for detecting psycho-emotional state has become a pressing need, which makes it possible to monitor the situation in real time and timely take the necessary measures to prevent the threat or eliminate the consequences of the incident. Defining aggression on video is a complex task, as the very concept of aggression definition can be vague and uncertain.

Every third schoolchild in the Republic of Kazakhstan has faced bullying and humiliation. These are the data of monitoring conducted in five schools in three regions of the country.

More than 50 % of surveyed children answered that there are facts of bullying in school and class. 33.6 % of pupils claim to have been victims of cyberbullying. Most often they are girls who received insulting messages by phone, via messengers or social networks [1]. Almost half of children admitted that high school students most often resort to bullying, with peers coming in second place. 10.3 % of respondents named teachers as initiators of bullying, and 5.6 % – parents, siblings. Students in grades 4–7 are more susceptible to bullying by older students.

National Center for Educational Statistics (2021) reports that one in five students (20.2 %) report bullying that occurred in places such as the hallway or stairwell of schools (43 %), cafeteria (27 %), school grounds (22 %), online/text (15 %), restroom or locker room (12 %), and school bus (8 %) [2].

The consequences of bullying are serious for both the victim and the bully. Victims usually find that they suffer from depression, physical illness and dissatisfaction with life for up to three years, and these factors in turn reduce their ability to enroll in universities and post-secondary institutions [3].

A Canadian study examined the link between teen suicide and bullying and sexual violence. Repeated online and real-life violence led to symptoms of PTSD and inadequate maternal support, leading to positive suicidal ideation among girls aged 12 to 18, and sexual harassment had nothing to do with suicide [4].

Similarly, the association between female gender and suicidal ideation was confirmed in a Chinese study involving 23,392 schoolchildren that tested the relationship between aggressive behavior and the influence of gender on suicidal thoughts and actions. Victims (sometimes bullies) were at higher risk of suicidal thoughts than neutral people. A similar situation can be found in the relationship between aggressive behavior and suicide attempts. Further hierarchical analysis showed that feelings of being bullied by others, suicidal thoughts and suicide attempts were slightly stronger in girls than in boys [5].

Unlike shallow models, most deep models do not require extracting individual features because they are based on a feature learning method that learns their features from the data and classifies them based on them. Moreover, in addition to completing the training, the extracted features can be given as input to SVMs and other shallow model classifiers [6].

2. Literature data analysis and problem statement

In [7], results of automatic detection of various abnormal behaviors in video clips are presented to solve the problem of suspicious action detection. It is shown that the system includes three main stages: moving object detection, object tracking and behavior understanding for action recognition. In the first pre-processing stage, moving objects are detected and noise is removed. This object extraction process has been used to detect basic features such as direction, velocity, size and centroid. However, the extracted features although help to track objects in fight clips, but do not detect emotions.

In [8], a study of crowd violence control is presented. Two related but different tasks are shown: violence classification and violence detection. The proposed approach does not take into account the statistics of changes in the magnitude of flow vectors over time.

The deep learning paradigm is applicable to a task that takes as input a complete video sequence using 3D CNNs. In [9], a hybrid framework" handcrafted/learned" is proposed. It is shown that the method aims, firstly, to obtain an illustrative image from the video sequence taken as input data for label retrieval. However, human motion features are very important for this task, and using the full image as input creates noise and redundancy in the learning process. All this suggests that the dual-threaded architecture may not be suitable for real-time applications due to computational complexity.

Body movement is the main interface on which more attention has been paid in recent years.

In surveillance systems, detection of suspicious activities plays a vital role. In many human activities, there is an urgent need for a surveillance system that can function reliably. That is why a new framework is proposed for real time surveillance system that detects abnormal human activity automatically [10].

A three-stage framework for recognizing violence is proposed for detecting violence in deep learning [11]. The work is divided into three phases. It is shown that the preprocessing phase includes human abnormal activity detection and content-based image extraction phase. But there are still outstanding issues related to the inclusion of body pose. This approach has been used to detect abnormal activity of students but does not utilize pose estimation.

In [12], a temporal differentiation algorithm is presented which is used to detect moving objects and then motion regions are detected using Gaussian function. It is shown that a shape model based on omega equation is used as a filter for the recognized objects which are human or non-human. Human activities are categorized as normal and abnormal activities using SVM. But there are still outstanding issues related to the case of abnormal activities. This approach has been used in software modeling using MATLAB and experimental results show that the system achieves tracking, semantic scene learning and anomaly detection in the environment without human involvement. However, the system does not work in crowded places.

In [13], a framework based on late fusion is proposed to analyze high-level activity using multi-independent temporal perceptual layers. Results obtained from multitemporal perception layers are presented. A framework for detecting violent events in visual observation is shown to evaluate this approach. Three well-known 29 datasets are used for the experiments: the NUS-HGA, Behavior and some YouTube videos. Experimental results show that using a multi-temporal method has no particular advantage over simultaneous methods.

In [14], violence videos are represented by combining statistical features. A model based on CNN BI feeds and SVM for violence detection is shown. The proposed model consists of three parts: feature extraction, SVM training and label fusion. Experiments are conducted to evaluate the proposed method using hockey fight and violent crowd datasets. Experimental results show that the proposed method performs better than existing methods like HOG, HOF, MoSIFT, SIFT, Two-stream in many realistic scenes in terms of accuracy. However, this method does not highlight the emotional actions used for violence detection.

In [15], state-of-the-art methods for 3d reconstruction of human poses are investigated. However, the tasks present several challenges: large dataset with long video clips, large number of highly variable actions. The age difference and unpredictable behaviour of autistic children further complicated the task. Models with these features show limited classification accuracy.

Recognizing and detecting violence based on image processing is becoming an important topic for outdoor surveillance systems. The main goal is to determine if violence is occurring.

In [16], local objects and their spatio-temporal positions are used to visualize the image. It is shown that the data structure of the area summary table is used to speed up the approach and the IFV formulas are rearranged. For this work, it is recommended to extend the improved Fisher vectors (IFV) for video clips. Experimental results show that the proposed method shows not high accuracy.

Based on the above sources, in the works despite the variety of approaches to learning and classification feature extraction most of the methods do not address the problem of recognizing violent movements under the action of invariant 3D transformations. Also, these models may not be suitable for real-time applications due to computational complexity.

The main problem is the lack of common solutions for tracking aggressive behavior and possible bullying in educational institutions. Therefore, there is a need to develop an AI model for recognizing objects on video, which will reduce computational costs and will be characterized by high accuracy and short training time.

3. The aim and objectives of the research

The aim of this study is to develop solutions based on artificial intelligence (AI) and create on their basis a prototype of a software and hardware complex capable of automatically detecting facts of aggressive behavior and possible violence and intimidation in educational institutions.

To achieve this aim, the following objectives are accomplished:

– collect and process raw video data from the Internet and schools in Almaty;

– develop machine learning models to provide automated detection of aggressive behavior and bullying;

– develop and test a prototype system to automatically identify potential physical bullying.

4. Materials and Methods of Research

The object of the study is the methods of video image processing to determine the psycho-emotional state of the observed using artificial intelligence.

Hypothesis of the study is that recognition of human emotions on the basis of video surveillance analysis with the use of artificial intelligence technology is an effective method in determining the aggressive state of the observed.

Training artificial intelligence models requires video markup of 3 types:

– temporal markings of the beginning and end of the studied incidents of potential bullying and/or aggressive behavior;

– spatial marks (rectangles) of people in the frame during the incidents under study;

– spatial markers (rectangles) of people's faces in the frame during the incidents under study.

According to the classifier used, violence detection methods are classified into three categories: violence detection methods using machine learning, violence detection methods using SVM, and violence detection methods using deep learning. SVM and deep learning are categorized separately because these algorithms are widely used in the field of computer vision. Deep learning has become a very popular area of machine learning, surpassing traditional methods in many computer vision applications. A very useful attribute of deep learning algorithms is the ability to extract features from raw data, eliminating the need for manual descriptors.

Due to the lack of consistency, attack events are difficult to define and usually require high-level interpretation.

Therefore, it is common to classify what is often present in videos of violent human behavior at a low level, i. e., unstructured and jerky movements. To achieve this goal, a PoseNet model is proposed.

Table 1 shows the list of methods for recognizing violent events. SVM (Support vector machine) is an algorithm used to solve classification tasks using reinforcement exercises. The classification task involves learning with a teacher. SVM algorithm is an algorithm for learning with a teacher. The main goal of SVM as a classifier is to find the equation of a divisible hypersurface in a space in the space R^n (1)–(3) which somehow optimally distinguishes between the two classes:

$$\omega_1 x_1 + \omega_2 x_2 + \ldots + \omega_n x_n + \omega_0 = 0, \tag{1}$$

where x is an object from the space R^n.

General view of converting F of an object to a class label Y:

$$F(x) = \operatorname{sign}\left(\omega^T x - b\right), \tag{2}$$

where ω and b are the weights of the algorithm and (learning).

It is possible to denote $\omega = \left(\omega_1, 2\omega_2, \ldots., \omega_n \omega n\right),\ b = -0\omega_0$.

The margin of an object x from the boundary of classes is the value $M = y\left(\omega^T x - b\right)$. The algorithm makes an error on an object if and only if the margin M is negative (when y and $(\omega^T x - b)$ are of different signs).

The algorithm correctly classifies objects if the following condition is satisfied:

$$y\left(\omega^T x - b\right) \geq 1.$$

If to combine the two derived expressions, let's get the default setting of hard-margin SVM, where no object is allowed to enter the split lane. The inequality is solved analytically using the Kuhn-Tucker theorem. The resulting problem is equal to the dual problem of finding the saddle point of the Lagrange function:

$$\frac{\left(\omega^T \omega\right)}{2} \to \min, y\left(\omega^T x - b\right) \geq 1. \tag{3}$$

To address the challenging problem of detecting violence from video, the PoseNet model was utilized. Was selected a lightweight pre-trained neural network, PoseNET, to identify human figures in the frame. The pretrained PoseNET network can be used inside an extractor function, which allows to transfer knowledge about frame features from one semantic space to another. The result of PoseNET is a representation of a human figure with coordinates and confidence estimates of 17 key points of the figure.

PoseNet can be used to estimate single or multiple poses, meaning that one version of the algorithm can identify only one person in an image/video, while another version can identify more than one person in an image/video.

Identification of a person:

– skeletons of human figures are identified in the video using the PoseNet model;

– the coordinates of the salient points are then used as landmarks for the aggression class classifier.

Table 1

Methods for detecting violence using SVM

Method	Method “Object detection”	Method “Exception extraction”	Method “Event type”	Accuracy %
Real-time violence detection in crowded places [16]	VIF descriptor	Feature set	Crowded	88 %
«Bag of words» framework with acceleration for action detection [17]	Background subtraction algorithms	Ellipse estimation method for consecutive frames	Less crowded	Approximately 90 %
Framework a genetic algorithm with a tracking and detection module [18]	Gaussian model	Algorithm for optical flow extraction	Crowded	82–89 %
Multimodal features of an internal class-based framework [19]	Image CNN and ImageNet	Google Net for feature extraction	Less crowded	98 %
Determination of the frequency of forced assignments [9]	Spatial pyramids and grids for object	Detection Spatial and temporal grid methods for object extraction	Crowded	96–99 % using different datasets
Violence detection using directed violence flow [20]	Optical flow method	Combination of ViF and OviF Descriptor	Crowded	90 %
Combined AEI and HOG system for recognizing abnormal events in visual movements [21]	AEI method for subtracting	HOG background and spatio-temporal feature extraction methods	Both crowded and less crowded	94–95 %
The framework includes preprocessing, activity detection, and image extraction. This work identifies anomalous events and data [22]	Optical flow and time difference for object detection CBIR image extraction method	Gaussian function for video file analysis	Less crowded	97 %
Late integration method for perceptual time layers for high level activity detection. [23]	Motion vector method for identification from multiple cameras in two dimensions	SGT MtPL method	Less crowded	98 %
Two-channel convolutional neural network for real-time detection [24]	ImageNet for object detection	VGG-f model for feature extraction	Crowded	91–94 %
Solve the detection problem by segmenting the target by depth and clear format using Connect [25]	Motion detection and Trof	BoW model approach	Less crowded	96 %
Bag-of-words method using spatio-temporal method for detecting anomalies in video [26]	Representation	Of HOF and HOG segments and subsegments to obtain video frames	Less crowded	84–91 %
Using multiple cameras from 1 to N. [26]	3D convolution is used to obtain video frames, spatial information	Error backpropagation method	91 % in crowded places	91 % in crowded places
Deep location recognition architecture [27]	VGG VLAD image search	Method Error backpropagation method for feature extraction	87–96 % in crowded places	87–96 % in crowded places
Violent scenes using CNN and Deep sound capabilities [28]	MFB	CNN model	About 90 % in crowded places	About 90 % in crowded places
Violence image detection with ConvLSTM [29]	CNN with	ConvLSTM CNN model	About 97 % in crowded places	About 97 % in crowded places

The pose estimation model uses the camera-processed image as input and output of key points. The identified keypoints are indexed by pattern ID with reliability scores ranging from 0.0 to 1.0. The reliability score refers to the probability that the key point is there. The main points of this method are described below.

The key point confidence score is what determines the confidence in the accuracy of the predicted location of key points. This value ranges from 0.0 to 1.0. It can be used to hide key points where there is not enough confidence.

The location of the key point is the two-dimensional *X* and *Y* coordinates of the key points defined in the original input image.

The main problem is the part of the person’s assumed pose such as nose, right ear, left knee, right leg, etc. It has confidence scores for poses and key points. PoseNet can currently identify 17 key points.

The body joints identified by the posture assessment model are shown in Table 2.

Action detection model is one of the AI software models for detecting aggressive behavior/physical bullying in videos.

Table 2

Various body joints found by the PoseNet model

ID	Part
0	Nose
1	Left eye
2	Right eye
3	Left ear
4	Right ear
5	Left shoulder
6	Right shoulder
7	Left elbow
8	Right elbow
9	Left wrist
10	Right wrist
11	Left hip
12	Right hip
13	Left knee
14	Right knee
15	Left side
16	Right ankle

Table 3

Signs of aggressive behavior

No.	Pattern of behavior of the "aggressor"	Aggressive behavior class	Pattern of "victim" behavior	Physical Aggression Class	Additional parameters
1	Frequency of raising your hands up (it is not possible to determine)	Takes something forcefully from the victim	covers the face	Torso tilt forward, directed at the victim	Grabs
2	Torso tilt forward, facing the victim	Pointing/pointing a finger at	Head down	Head tilt forward, directed towards the victim	Throws an object
3	Head tilt forward, facing the victim	the victim	Lying down	Head raised high	Spits
4	Head held high	Slapping	Lifting bent legs up	Legs spread wide	Pours something
5	Legs spread wide	the back of the head Slapping the face	staggering,	Height of the elbow joint relative to the torso	Sprays in the face (gas can)
6	Stride width (wide stride towards the victim) (can't determine)	Waving and gesticulating in front of the victim	limping	Reduction of the aggressor's distance with the victim	Sets fire
7	Height of the elbow joint relative to the torso	Waving	Running away from the aggressor	Gesturing in a horizontal plane	Thrusts knife
8	Sudden movements (it is not possible to determine)	Headbutt	Reduced frequency of blinking eyes (it is not possible to determine)	Gesturing in a vertical plane	Fires weapon
9	Shortening the aggressor's distance from the victim	Handbutt	Avoiding eye contact (can't identify)	The toes of the feet point in different directions	Directs weapon (takes aim)
10	Gesturing in a horizontal plane	Kick	Looks around/looks	Position of hands on thighs	Suspends (takes it by the scruff of the neck)

5. Research results on determining the psycho-emotional state of the observed based on the analysis of video observations

5. 1. Collection and processing of raw video data

Written permission was obtained from the Almaty City Education Department to access the data from the school CCTV cameras.

Once the data were collected, they were prepared for training by the program module. The collected video images were reviewed by professional psychiatrists, and they conducted qualitative and quantitative analysis of the actions of participants of scenes with aggressive behavior.

Based on the analysis, a list of 186 preliminary parameters characteristic of scenes with aggressive behavior was developed. The formulated parameters were grouped into the following categories:

– "aggressor" behavior pattern;
– "aggressive behavior" class;
– "victim" behavior pattern;
– "physical aggression" class;
– additional parameters.

The list of preliminary parameters of behavior characteristic of scenes with aggressive behavior is given in Table 3.

Once the data were collected, they were prepared for training by a software module. The collected video images were reviewed by professional psychiatrists and they conducted qualitative and quantitative analysis of the actions of participants in scenes with aggressive behavior.

After generating the list of preliminary parameters, the markup of the collected videos was performed.

5. 2. Developing a machine learning model

To develop an AI model for recognizing objects in video, skeleton model training was used. This approach reduces the computational cost. A lightweight pre-trained neural network, PoseNET, was chosen to identify human figures in the frame.

The output of PoseNET is a representation of the human figure using 17 key points with coordinates and reliability scores. These 17 points include: nose, eyes, ears, upper arm, elbows, forearms, hips, knees, and ankles. An example of identifying the 17 key points via PoseNET is shown in Fig. 1.

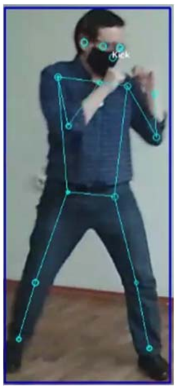


Fig. 1. 17 key points through the PoseNET network [10]

PoseNet returns the confidence value of each identified person and the key points of each identified gesture.

Starting from the top level, pose estimation is done in two steps:

1. An RGB image is inserted into a convolutional neural network.

2. Single or multi-pose decoding algorithms are used to decode poses, construct reliability estimates, locate key points and estimate the reliability of key points based on the model output.

A full-link layer of neural network was used to classify the obtained representation of human figure. This neural network superstructure is trained by minimizing a categorical cross-entropy loss function. The classification results are normalized by the next layer c multivariable logistic function (Softmax). The architecture of the used PoseNET neural network is shown in Fig. 2.

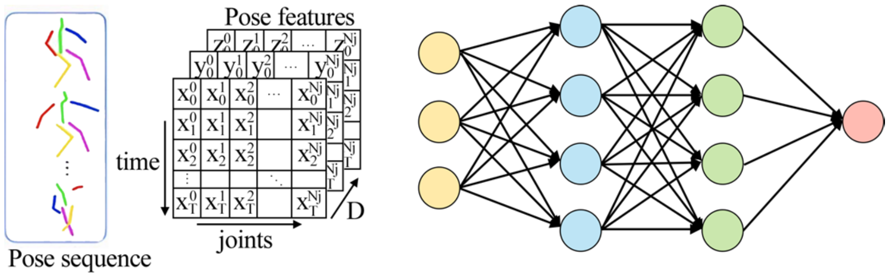


Fig. 2. Architecture of the used PoseNET neural network

The experiment of training the artificial intelligence model to recognise objects on video was conducted in accordance with a widely used scheme: the data set was divided into two parts – training and test – the proportions were 80 % and 20 % respectively.

As can be seen from Fig. 3, 4, after 8 training periods, the accuracy rate of the artificial intelligence model in identifying objects on video reached 98 %. Based on the results of numerous experiments, the model that showed the best quality was selected for identifying objects on video.

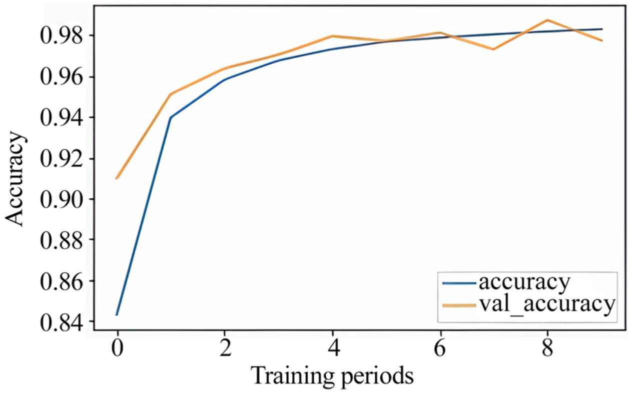


Fig. 3. Change in model accuracy during training

Fig. 3 shows the change in model accuracy during the training process. Fig. 4 shows the change in the loss function of the model during the training process.

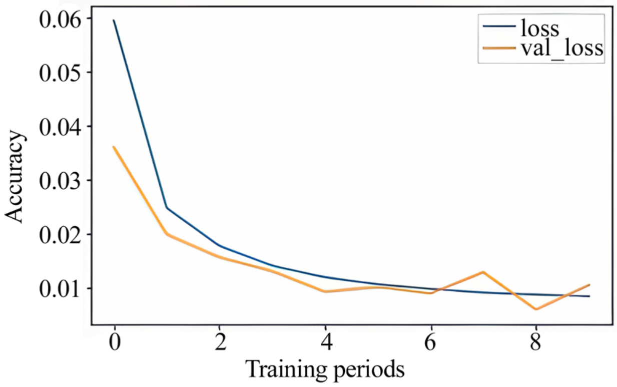
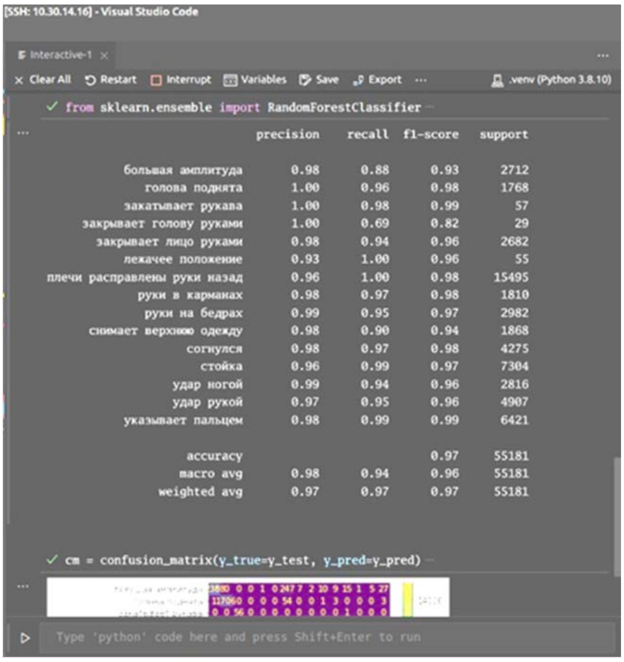


Fig. 4. Change in model loss function during the training process

5. 3. Development and testing of a prototype system

Based on the AI software model for recognizing objects in videos, an AI software model for tracking objects in videos was developed. For the experiments of tracking objects on video, 13 labels were selected from all the labels describing aggressive behavior. The AI model was then trained to track human figures in the video and classify their actions during the tracking process.

After the training of the AI software models to detect aggressive behavior/physical bullying in videos was completed, their quality was tested. The results of the testing are shown in Fig. 5.



```
from sklearn.ensemble import RandomForestClassifier

                           precision    recall  f1-score   support

        большая амплитуда       0.98      0.88      0.93      2712
           голова поднята       1.00      0.96      0.98      1768
         закатывает рукава      1.00      0.98      0.99        57
   закрывает голову руками      1.00      0.69      0.82        29
     закрывает лицо руками      0.98      0.94      0.96      2682
         лежачее положение      0.93      1.00      0.96        55
плечи расправлены руки назад    0.96      1.00      0.98     15495
           руки в карманах      0.98      0.97      0.98      1810
            руки на бедрах      0.99      0.95      0.97      2982
   снимает верхнюю одежду       0.98      0.90      0.94      1868
                  согнулся      0.98      0.97      0.98      4275
                    стойка      0.96      0.99      0.97      7304
                удар ногой      0.99      0.94      0.96      2816
                удар рукой      0.97      0.95      0.96      4907
         указывает пальцем      0.98      0.99      0.99      6421

                  accuracy                          0.97     55181
                 macro avg      0.98      0.94      0.96     55181
              weighted avg      0.97      0.97      0.97     55181

cm = confusion_matrix(y_true=y_test, y_pred=y_pred)

Type 'python' code here and press Shift+Enter to run
```

Fig. 5. Testing results of AI software models

As can be seen, the best results (f1 score with values of 0.98–0.99) are for the parameters “points with finger”, “rolls up sleeves”, “head raised”, “shoulders straightened with hands back”, “hands in pockets”, “bent”. The worst result was shown by the parameter “covers head with hands”, f1 score is equal to 0.82. The weighted average value of the f1-score metric was 0.96, which is quite a high value.

The results of the AI software model of object tracking on video can be represented as a confusion matrix. The classes assigned by the model are represented in the rows of the matrix, the true values of the classes are represented in the columns of the matrix. Table 4 shows the model’s learning quality metrics.

Table 4

Metrics of model training quality

Mark	Label Precision (precision)	Completeness (recall)	F1-measure (F1 – score)
Increased range of movement of the head arms and legs	0.98	0.97	0.98
Head held high	0.99	1.00	0.99
Torso tilt forward, directed at the victim	1.00	0.99	0.99
Straightened shoulders and arms back	0.98	0.99	0.99
Palms on the hips	0.98	0.99	0.99
Removes outer clothing	0.97	0.96	0.96
Kick	1.00	1.00	1.00
Hand punch	1.00	1.00	1.00
Covers the face with your hands	0.99	1.00	1.00
Toes pointing in different directions	0.99	1.00	1.00
Bouncing in place during a series of punches	0.99	1.00	1.00
Bent	over 1.00	0.98	0.99
Pointing Pointing at the victim	0.99	0.99	0.99
Micro averaging 0.99	0.99	0.99	0.99
Macro averaging 0.99	0.99	0,99	0.99
Weighted averaging 0.99	0.99	0.99	0.99

Fig. 6, 7 show examples of how an AI software model works for tracking objects in a video.

So, an artificial intelligence software model for tracking objects in video has been developed, which is characterized by high values of quality metrics: precision (precision), completeness (recall), F1-score (F1-score).

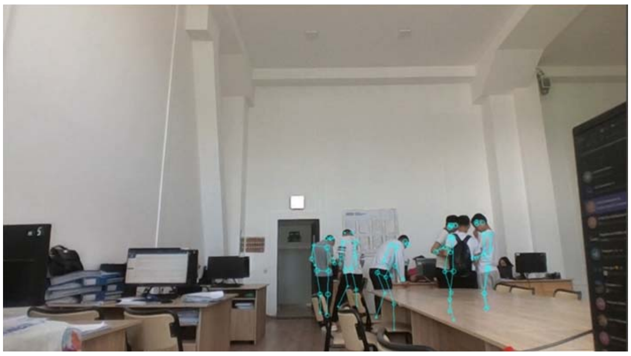


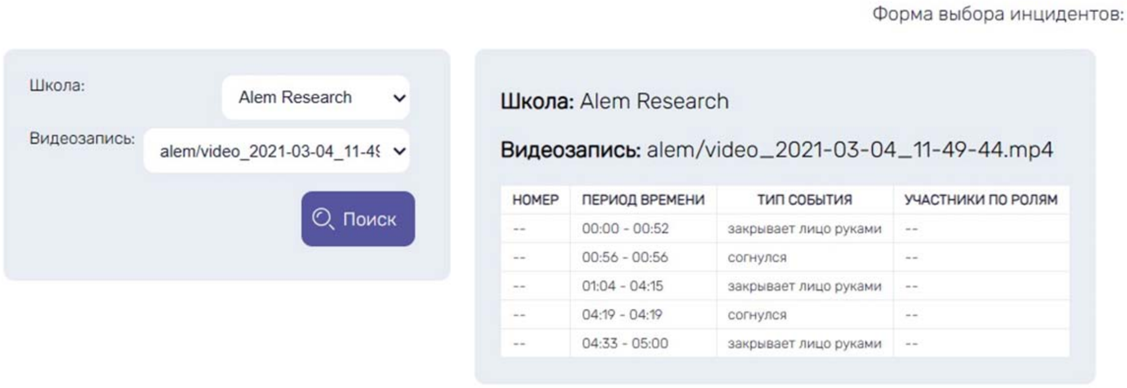
Fig. 7. Operation of the AI software model for tracking objects on video

6. Discussion of the results of developing an artificial intelligence model for detecting aggressive behavior

This paper analyzes machine learning and artificial intelligence algorithms for automatic detection and classification of physical, social and other types of violence, as well as for detecting the observed psycho-emotional state. A comparative analysis of these algorithms with examples of other works and studies of foreign specialists was carried out. The result of this project is a prototype of a software and hardware complex. Video content received from surveillance cameras located inside the building (in corridors, canteens, etc.) will be sequentially analyzed by the system in order to identify incidents and determine their type. When recording visual signals of physical aggression, the facial recognition system will memorize the participants of the incident, categorize them into roles (“initiator of violence” and “victim”) and record repeated incidents with their participation.

As can be seen from the figure (Fig. 3, 4), after 8 training periods, the accuracy rate of the AI model in identifying objects in the video reached 98 %.

The results of testing of program models AI showed the best results (f1 score with values 0.98–0.99) for the parameters “points finger”, “rolls up sleeves”, “head raised”, “shoulders straightened hands back”, “hands in pockets”, “bent”. The worst result was shown by the parameter “covers head with hands”, f1 score equals 0.82. The weighted average value of the f1-score metric was 0.96, which is quite a high value (Fig. 5).



Форма выбора инцидентов:

Школа: Alem Research

Видеозапись: alem/video_2021-03-04_11-4(

Школа: Alem Research

Видеозапись: alem/video_2021-03-04_11-49-44.mp4

НОМЕР	ПЕРИОД ВРЕМЕНИ	ТИП СОБЫТИЯ	УЧАСТНИКИ ПО РОЛЯМ
--	00:00 - 00:52	закрывает лицо руками	--
--	00:56 - 00:56	согнулся	--
--	01:04 - 04:15	закрывает лицо руками	--
--	04:19 - 04:19	согнулся	--
--	04:33 - 05:00	закрывает лицо руками	--

Fig. 6. Incident selection form

In addition, an attractive feature of the proposed method is that its performance is not affected by the number of people in the input image regardless of whether it is 15 people or 5 people, the computation time is the same. The pre-trained PoseNET can be used inside the feature extractor, which allows the knowledge of frame features to be transferred from one semantic space to another. Which is not the case for the other works reviewed here. For example, in [12] the obtained features help to track objects in clips with battles, but do not determine the psycho-emotional state of players.

In the tracking system, detection of suspicious activities plays an important role. Researchers urgently need a monitoring system that can work reliably. Therefore, this paper proposes a new structure of real-time video surveillance system with automatic detection function.

The used artificial intelligence model was trained to track human figures in the video and classify their actions during the tracking process.

The conducted research of the system has shown great efficiency of the method of determining changes in the emotional state of a person on the basis of video surveillance. The advantage of this system is high accuracy of its work (97 %) (Fig. 5).

The video materials to be studied must be divided into parts up to 1 hour in length with each part assigned a unique identifier. Each part must be viewed at least 3 times in order to perform one of the markup types. It is not recommended to perform different types of markups at the same time due to quality requirements.

It should also be noted that sometimes false positives may occur for actions of people who were not present in the training sample (for example: a technician removing rubbish from the floor in a sitting position may be labelled by the prototype system as “lying down”; a schoolboy fixing his hair may be labelled by the prototype system as “covering his face with his hands”).

The reason for these deficiencies is the peculiarities of the data on which the AI software models were trained. Adding to the training sample video recordings with situations that caused false responses of AI software models corrected the identified defects. In the future, in order to minimize the number of false positives of the system, in each school where it will be used, at the first stage of its implementation it is necessary to carry out additional training of AI software models of the system on the available actual video recordings.

7. Conclusions

1. Raw video data from the Internet and schools of Almaty city were collected for training AI program models. After forming the list of preliminary parameters, the markup of the collected videos was performed. As a result, a dataset with the following characteristics was obtained:

– number of labeled segments: 21 237;

– total duration of the marked-up segments: 12 hours 26 minutes 10 seconds.

2. The artificial intelligence software model for recognizing objects in video based on PoseNET architecture has been developed, which is characterized by high accuracy and short training time. This can be seen from the result obtained by testing artificial intelligence program models. A special feature is the use of PoseNet, implemented using PyTorch, which can accurately detect the pose of a person in a video and extract information about the position of different body parts.

3. A prototype of an artificial intelligence-based software solution for automatic detection of potential physical bullying was developed. The developed machine learning models for detecting aggressive behavior in videos were tested and showed high results-0.98. This is enabled by the use of technical and software tools such as PoseNet, PyTorch, SQLite, FastAPI and Flask, which presents significant potential and many possibilities.

References

1. Bauman, S. (2016). Do We Need More Measures of Bullying? Journal of Adolescent Health, 59 (5), 487–488. https://doi.org/10.1016/j.jadohealth.2016.08.021
2. Al-Nawashi, M., Al-Hazaimeh, O. M., Saraee, M. (2016). A novel framework for intelligent surveillance system based on abnormal human activity detection in academic environments. Neural Computing and Applications, 28 (S1), 565–572. https://doi.org/10.1007/s00521-016-2363-z
3. Seldin, M., Yanez, C. (2019). Student Reports of Bullying: Results from the 2017 School Crime Supplement to the National Crime Victimization Survey. Web Tables. NCES 2019–054. National Center for Education Statistics. Available at: https://nces.ed.gov/pubs2019/2019054.pdf
4. McCarthy, R. J., Elson, M. (2018). A Conceptual Review of Lab-Based Aggression Paradigms. Collabra: Psychology, 4 (1). https://doi.org/10.1525/collabra.104
5. Parrott, D. J., Zeichner, A. (2002). Effects of alcohol and trait anger on physical aggression in men. Journal of Studies on Alcohol, 63 (2), 196–204. https://doi.org/10.15288/jsa.2002.63.196
6. Allen, J. J., Anderson, C. A. (2017). Aggression and Violence: Definitions and Distinctions. The Wiley Handbook of Violence and Aggression, 1–14. https://doi.org/10.1002/9781119057574.whbva001
7. Zhou, P., Ding, Q., Luo, H., Hou, X. (2018). Violence detection in surveillance video using low-level features. PLOS ONE, 13 (10), e0203668. https://doi.org/10.1371/journal.pone.0203668
8. Hassner, T., Itcher, Y., Kliper-Gross, O. (2012). Violent flows: Real-time detection of violent crowd behavior. 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. https://doi.org/10.1109/cvprw.2012.6239348
9. Ullah, F. U. M., Ullah, A., Muhammad, K., Haq, I. U., Baik, S. W. (2019). Violence Detection Using Spatiotemporal Features with 3D Convolutional Neural Network. Sensors, 19 (11), 2472. https://doi.org/10.3390/s19112472
10. Amirgaliyev, Y. N., Bukenova, I. N., Bukenov, G. S., Kenshimov, C. A. (2023). Software solutions for the recognition violent movements by video. Bulletin of East Kazakhstan Technical University, 2, 31–42.

11. Peixoto, B. M., Avila, S., Dias, Z., Rocha, A. (2018). Breaking down violence. Proceedings of the 13th International Conference on Availability, Reliability and Security. https://doi.org/10.1145/3230833.3232809
12. Song, D., Kim, C., Park, S.-K. (2018). A multi-temporal framework for high-level activity analysis: Violent event detection in visual surveillance. Information Sciences, 447, 83–103. https://doi.org/10.1016/j.ins.2018.02.065
13. Carneiro, S. A., da Silva, G. P., Guimaraes, S. J. F., Pedrini, H. (2019). Fight Detection in Video Sequences Based on Multi-Stream Convolutional Neural Networks. 2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). https://doi.org/10.1109/sibgrapi.2019.00010
14. Febin, I. P., Jayasree, K., Joy, P. T. (2019). Violence detection in videos for an intelligent surveillance system using MoBSIFT and movement filtering algorithm. Pattern Analysis and Applications, 23 (2), 611–623. https://doi.org/10.1007/s10044-019-00821-3
15. Marinoiu, E., Zanfir, M., Olaru, V., Sminchisescu, C. (2018). 3D Human Sensing, Action and Emotion Recognition in Robot Assisted Therapy of Children with Autism. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/cvpr.2018.00230
16. Koppula, H. S., Gupta, R., Saxena, A. (2013). Learning human activities and object affordances from RGB-D videos. The International Journal of Robotics Research, 32 (8), 951–970. https://doi.org/10.1177/0278364913478446
17. Bermejo Nievas, E., Deniz Suarez, O., Bueno García, G., Sukthankar, R. (2011). Violence Detection in Video Using Computer Vision Techniques. Lecture Notes in Computer Science, 332–339. https://doi.org/10.1007/978-3-642-23678-5_39
18. Zhou, P., Ding, Q., Luo, H., Hou, X. (2017). Violent Interaction Detection in Video Based on Deep Learning. Journal of Physics: Conference Series, 844, 012044. https://doi.org/10.1088/1742-6596/844/1/012044
19. Pawar, K., Attar, V. (2018). Deep learning approaches for video-based anomalous activity detection. World Wide Web, 22 (2), 571–601. https://doi.org/10.1007/s11280-018-0582-1
20. Zhao, H., Torralba, A., Torresani, L., Yan, Z. (2019). HACS: Human Action Clips and Segments Dataset for Recognition and Temporal Localization. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). https://doi.org/10.1109/iccv.2019.00876
21. Olweus, D. (1978). Aggression in the schools: Bullies and whipping boys. Hemisphere.
22. Solberg, M. E., Olweus, D. (2003). Prevalence estimation of school bullying with the Olweus Bully/Victim Questionnaire. Aggressive Behavior, 29 (3), 239–268. https://doi.org/10.1002/ab.10047
23. Shetgiri, R. (2013). Bullying and Victimization Among Children. Advances in Pediatrics, 60 (1), 33–51. https://doi.org/10.1016/j.yapd.2013.04.004
24. Fung, A. L. C. (2019). Adolescent Reactive and Proactive Aggression, and Bullying in Hong Kong: Prevalence, Psychosocial Correlates, and Prevention. Journal of Adolescent Health, 64 (6), S65–S72. https://doi.org/10.1016/j.jadohealth.2018.09.018
25. Lereya, S. T., Copeland, W. E., Costello, E. J., Wolke, D. (2015). Adult mental health consequences of peer bullying and maltreatment in childhood: two cohorts in two countries. The Lancet Psychiatry, 2 (6), 524–531. https://doi.org/10.1016/s2215-0366(15)00165-0
26. Buch-Frohlich, A., Paradis, A., Hébert, M., Cyr, M., Frappier, J.-Y. (2019). Bullying and sexual harassment as predictors of suicidality in sexually abused adolescent girls. International Journal of Victimology, 35, 63–73.
27. Yang, T., Guo, L., Hong, F., Wang, Z., Yu, Y., Lu, C. (2020). Association Between Bullying and Suicidal Behavior Among Chinese Adolescents: An Analysis of Gender Differences. Psychology Research and Behavior Management, Volume 13, 89–96. https://doi.org/10.2147/prbm.s228007
28. Lloyd, K., Rosin, P. L., Marshall, D., Moore, S. C. (2017). Detecting violent and abnormal crowd activity using temporal analysis of grey level co-occurrence matrix (GLCM)-based texture measures. Machine Vision and Applications, 28 (3-4), 361–371. https://doi.org/10.1007/s00138-017-0830-x
29. Bilinski, P., Bremond, F. (2016). Human violence recognition and detection in surveillance videos. 2016 13th IEEE International Conference on Advanced Video and Signal Based