information

Article

A Shallow System Prototype for Violent Action Detection in Italian Public Schools

Erica Perseghin * and Gian Luca Foresti *

Department of Mathematics, Computer Science and Physics, University of Udine, 33100 Udine, Italy
* Correspondence: perseghin.erica@uniud.it (E.P.); gianluca.foresti@uniud.it (G.L.F.)

Abstract: This paper presents a novel low-cost integrated system prototype, called School Violence Detection system (SVD), based on a 2D Convolutional Neural Network (CNN). It is used for classifying and identifying automatically violent actions in educational environments based on shallow cost hardware. Moreover, the paper fills the gap of real datasets in educational environments by proposing a new one, called Daily School Break dataset (DSB), containing original videos recorded in an Italian high school yard. The proposed CNN has been pre-trained with an ImageNet model and a transfer learning approach. To extend its capabilities, the DSB was enriched with online images representing students in school environments. Experimental results analyze the classification performances of the SVD and investigate how it performs through the proposed DSB dataset. The SVD, which achieves a recognition accuracy of 95%, is considered computably efficient and low-cost. It could be adapted to other scenarios such as school arenas, gyms, playgrounds, etc.

Keywords: Violence Action Detection; CNN; transfer learning; deep learning

check for updates

Citation: Perseghin, E.; Foresti, G.L. A Shallow System Prototype for Violent Action Detection in Italian Public Schools. *Information* **2023**, *14*, 240. https://doi.org/10.3390/info14040240

Academic Editors: Marco Leo and Sara Colantonio

Received: 24 February 2023
Revised: 26 March 2023
Accepted: 8 April 2023
Published: 14 April 2023

1. Introduction

Due to the recent increase in applications and systems based on action recognition, automatic classification of images and videos is becoming a key issue. Over the last decades, there has been a general increase in scenes of violence in public environments, especially generated by young people. Areas of particular interest are public schools; therefore, the using of surveillance camera systems could help to automatically trace and prevent violence. Unfortunately, as these data are not always collectable, there is a lack of available public datasets in school environments.

In the past ten years, the developing of specific digitalization plans started in Italian public schools, but reluctance, lack of knowledge in digital issues and in technical specialists limit the growth and the use of digital platforms. Investing in technology today is still expensive and it requires a complete digital transformation. It is an ambitious plan that still today requires time and homogeneity all across the Italian territory. Nowadays, several schools in Italy have old technologies and they do not often use CCTV systems due to privacy regulations. Thus, in this scenario, besides the technological improvements that have made possible sophisticated detection techniques, there is the need of making intelligent and autonomous violence detection systems based on public school shallow cost technologies. Indeed, higher computational complexity requires high performance computing in order to train CPU-based implementations which are not available in public schools.

Along this road, one of the first things is to explore the different possibilities of automatic violence detection using artificial intelligence.

Convolutional Neural Networks (CNNs) are nowadays dominating over visual recognition tasks, providing excellent results in many research fields [1,2]. Despite the big success of CNNs, a limitation is represented by the datasets for training purposes. Datasets are a fundamental part of machine learning approaches.

In literature, there are few datasets specific to violence recognition, but there are none for violence detection in educational environments. Generally, input images/videos are

in low resolution [3], within specific contexts [4] or re-create fake scenes (using actors, no occlusions on stage, no noises, etc.) [5]. Existing datasets contain few frames due to the nature of the recorded images: actions, movements, lights, shadows, indoor/outdoor spaces, etc., do not always allow for easy recognition of the action. In a nutshell, datasets described in the state-of-art literature achieved good performances in detecting violence, but they do not reflect the reality. Usually, they are formed by violent actions in sports, movies or extracted from online videos. Specifically, there is no current study that has considered the creation of a dataset containing real scenes of an authentic school environment, except for a few pioneer papers [5–7] that have re-created indoor sets with a small number of acting students.

In this setting, the main innovative content of the paper is represented by:

1. The design and the development of a School Violence Detection (SVD) system, a low-cost integrated system, allowing for easy binary violence detection. It is consolidated by transfer learning techniques in order to extend its potentialities.
2. The creation of Daily School Break (DSB), a new dataset of images and videos of Italian students recorded in school environments. Original for typology and subjects, this kind of dataset does not exist in the literature, and it can be used to compare different techniques.

The SVD prototype can open the door to various possibilities. It could end up being used in schools, public offices, parishes, etc. It is cheaper, safer and easier to change and iterate over again as the system does not request advanced hardware technologies and the processing time is affordable. It can respond with a quick reply, alerting when it recognizes violence actions.

DSB is a labeled dataset for supervised and semi-supervised machine learning algorithms. The videos show daily school break scenes in a school yard in which the students are free to move around the area without any restrictions, performing typical activities such as eating, chatting, walking, etc. They include also possible violent scenes such as kicks, hits, or pushes. In detail, the DSB dataset was manually labelled in a binary way: violent and non-violent images. No filtering or modifications were applied to the scenes in order to really analyze how a machine learning approach is able to catch violence images for detecting violent actions in scenes of everyday school life.

The paper is organized as follows. In Section 2, the current state of the art on violence detection systems in real environments and datasets for violence detection purposes are analyzed and discussed. Section 3 presents the Daily School Break (DSB) dataset, describing the type and the content of the recorded videos. Section 4 shows the proposed School Violence Detection (SVD) system based on a 2DCNN architecture for binary classification of violent or non-violent scenes. Section 5 reports the experimental results of the proposed SVD applied to the images of the DSB dataset. Finally, Section 6 concludes the paper and highlights future works.

2. Related Works

Violence detection is becoming an important matter, not only for systems in the surveillance domain, but also in a range of other areas such as airports, metro stations, public spaces, prisons and hospitals. The state of the art of the last years contains several methods for human action recognition [8–22], despite violence detection and relative datasets still remaining a challenging task due to the presence of complex patterns and sequential information. In contrast to previous research, there is a relatively small amount of papers that investigate violence detection among young people: it is extremely rare to find studies into educational environments. Moreover, compared to single user activity recognition, the identification of interactions between two or more young people is less popular among researchers and requires further investigation. The recognition of the human body (such as postures or gestures) can introduce also complex social and psychological aspects, which depend on people's feelings influenced by context, culture and personal attitudes. The research methods with wearable devices [23] cannot carry out contactless action recognition

and it is difficult to use them in detecting young violence. Instead, deep learning is adopted for action recognition because it can automatically learn intrinsic features from images and output classification results. Regarding deep learning-based methods, as demonstrated in [24], the recent interest in violence detection has increased the study of several aspects such as in CNN, RNN, GCN, etc. Many of them can be integrated (2DCNN with 3DCNN in [25]) or combined with other deep models (as LSTM in [26], which involves offline model training and online testing phases).

In the next two subsections, previous works in the educational area and related datasets for violence detection are summarized.

2.1. Violence Detection Systems

A pioneer paper investigated violence detection in a campus is presented in [6]. The authors proposed a method of action recognition based on skeleton information [27,28] extracted from campus surveillance videos. A graph convolutional network (GCN) was used to analyze the information about posing, processing skeleton data and classification score in order to identify violent actions. In detail, in the selected dataset, 60 analyzed actions were single-person actions while other images show only two characters (student actors) without any interference in front of the camera. Although the method proposed in [6] reached a higher accuracy level, the chosen dataset for training and validation suffers different limitations, including the presence of specific and selected violent classes such as pushing, punching and kicking. These actions are widespread in schools, but they are not the only ones. At the present time, in reality, it is not possible to have perfect light conditions, noise or absence of occlusions in front of the camera.

The main goal of violence detection is the prevention, detection and possible intervention to contrast the violence. Starting from this idea, we ended up developing a real and consistent video dataset, which aimed to detect violence in a public environment without removing the background. We deliberately decided not to use any publicly available datasets as many papers in the-state-of-art do, due to fake violence action scenes or specific scenes. We created a new dataset in a still not explored context, such as the public high school yard. In the proposed approach, similar to [11], we implemented a transfer learning method based on ImageNet [29] in order to improve the classification by pre-trained model layers.

An interesting modus operandi for violence detection is to use a 2DCNN with additional convolutional layers for a faster encoding [11]. In [11], authors explored different strategies to extract salient features applying three ImageNet pre-trained models (VGG16, VGG19 and ResNet50) as summarized in Table 1.

Table 1. Main characteristics of ImageNet pre-trained networks.

Model	Year	Parameters	Layers
VGG-16	2014	138 million	16 weighted layers (Figure 1)
VGG-19	2014	143 million	2 more convolutional layers than VVG16
ResNet-50	2015	25 million	48 convolutional layers, 1 max-pooling and an average pooling. Its architecture is based on residual learning (Figure 2)

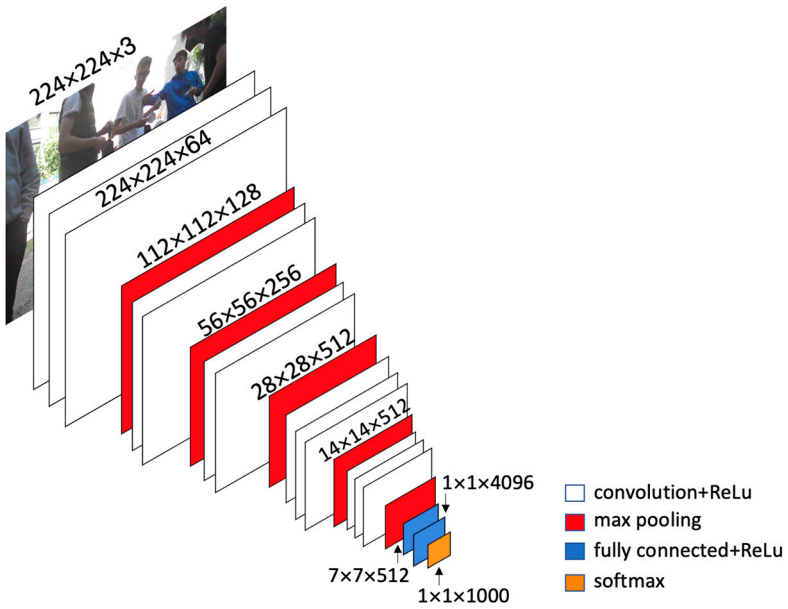


Figure 1. In VGG16 architecture [9], at the initial stage, the image is processed by a stack of convolutional layers of 3 × 3 filters and then a max pool layer of 2 × 2 filters. All hidden layers have been used with rectified Linear Unit (ReLu) being the activation function. The architecture follows the same structure until the end where it has 2 fully connected layers and a softmax for output.

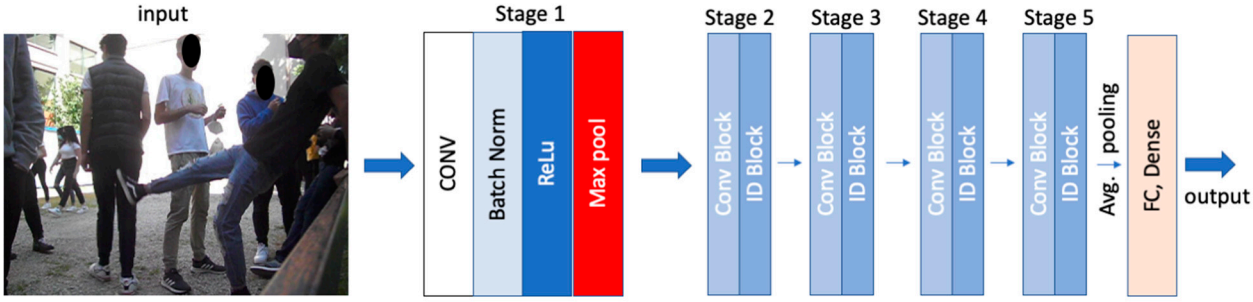


Figure 2. ResNet-50 architecture is a Convolutional Neural Network with 50 deep layers. It uses a bottleneck design: convolutions reduce the number of parameters and matrix multiplications enable faster training of each layer. Every convolutional layer is followed by a batch normalization layer, a ReLu activation function and a max pooling layer. Then, ResNet block transforms the input into the desired shape for the addition operation passing it through the next stages (residual blocks). At the final stage, average pooling followed by a fully connected layer works with a softmax activation function.

We summarize the points of particular interest in the following:

- VGG16 is one of the most famous CNN architectures as it has been awarded the ILSVR ImageNet competition in 2014 [9]. VGG16, instead of having a large number of hyper-parameters, is focalized on the convolution layer. Its architecture combines layers of 3 × 3 filter of stride one and a maxpool layer of 2 × 2 filter of stride two.
- VGG19 is similar to the previous one, but deeper than VGG16. It can better recognize the image in terms of color and structure. The architecture of the fully connected layers is the same in all VGG networks, making this model quite easy to implement.
- ResNet-50 is a deep residual network and is a more accurate CNN subclass. It is organized in residual blocks. It adopts a technique called skip connection that allows the knowledge of processed data to be transfered to the next level by adding the output of the previous levels. In shallow neural networks, consecutive hidden layers are linked to each other, but in ResNet architecture, information is passed from the down

sampling layers to the up sampling layers. The connection speeds up the training time and increases the capacity of the network [30]. Ref. [31] analyzed how ResNet-50 learns better color and texture features during the training phase due to the residual connection that avoids information loss [32].

Despite the state-of-the-art in violence detection techniques, applying complex CNNs can be difficult: they are limited to model geometric transformation, variations in pose or viewpoints, etc., but they are the most affordable and lightweight solutions. Moreover, in real scenarios, shallow and old infrastructures in schools can reduce the possibilities to work correctly. In addition, deeper approaches bring rapid growth in time and in energy consumptions that are not affordable for Italian public architectures. In the light of such considerations, we opted for a low-cost, small and agile system in order to have a rather high degree of detection. Moreover, the proposed SVD is implemented in shallow technologies, normally used in Italian schools. The proposed method presents a low detection time and tries to detect violence actions between close people in cloud scenarios where resolution, lightness or blurring can affect the detection process. Content-based analysis of multimedia regarding violence detection is numerous, but as analyzed in [24], the most common sources of datasets are taken from Hollywood movies, action movies or mobile phones. Differentially, we provided an original dataset of images of high school students which can be used as a common benchmark to compare different solutions in this field.

2.2. Violence Detection Datasets

The following is a concise review of the existing violence detection datasets:

- Hockey Fight [11]: this contains 1000 clips divided into two groups of 500 labelled sequences (fight and non-fight for binary classification) extracted from the National Hockey League (NHL). Each clip consists of 50 low resolution frames. The main disadvantage is the lack of the diversity because the videos are captured in a single scene. Moreover, [24] presented the complexity of detecting these types of indoor scene.
- Movie Fight [33]: this consists of 200 video clips in which the scenes of violence are extracted from action movies. As with the previous one, both of these two datasets have video-level annotations, but nowadays the last one is less used due to the insufficient number of videos. In addition, all activities in [11,33] are captured in the center of the view.
- Dataset for Automatic Violence Detection [5] was co-sponsored by the Italian "Gabinetto Interregionale di Polizia Scientifica per le Marche e l'Abruzzo". The aim of the project was to create a dataset with high resolution images describing violent situations. The clips were recorded in indoor environments with non-professional actors and the dataset is composed of 350 clips; 120 clips present non-violent behaviors and 230 clips violent behaviors. In particular, non-violent clips include hand claps, hugs, scenes of exultation, etc., which can cause false positives within a violence action recognition task due to the rapidity of the movements and the similarity to violent behaviors. The purpose of the dataset is to improve performance within the techniques of violence recognition and to verify their robustness. The violent clips include kicking, slapping, choking, gun shooting, etc. All the videos, tagged manually, were recorded inside the same room in natural light conditions. The primary limitation of this dataset is that the settings and lighting do not accurately reflect real-world environments.
- UCF-Crime [12]: this contains 1900 uncut or unedited videos recorded by surveillance cameras from all over the world. It is a large-scale dataset with over 128 hours of videos downloaded from YouTube via keyword research in multiple languages. It is used for violence classification and violence detection tasks as it provides a wide range of challenging real-world data within in-the-wild conditions. Darkness scenes, long length of videos with few level annotations and anomalous videos generate false alarms in the tested model [12]. Moreover, anomalies such as arrest, abuse, burglary,

explosion, etc. have a significant impact on public safety, but they cannot be compared to the public actions usually expected in school environments.

- RWF-2000 [3]: this is the largest dataset currently available. It includes 2000 videos extracted from surveillance cameras with real scenes captured all around the world. They were collected from YouTube via keywords (e.g., "real fights" or "violent events"). The authors developed a program that automatically downloads YouTube videos based on keyword research. They proposed a new method based on 3DCNNs [34]. They achieved a level of 87% accuracy. Unfortunately, sometimes rapid movements could be wrongly targeted, especially during the prevention of false positives [5]. Since all the videos are captured by surveillance cameras, many images may not have a good quality. As a consequence, only some of the involved people appeared in the pictures and some scenes are crowded and chaotic.

Several works [3,5,11,12,33] share commonalities, such as transient actions and specific contexts; these are not symbolic for detecting violence actions in educational environments. In addition, in [3], the design is customized for several operations, and it requests more cores to complete instructions quickly. On the other hand, in [12] there are overpopulated scenes that analyzed 13 realistic anomalies such as road accidents, burglaries, robberies, etc. Low image quality, lack of sufficient data, videos with long duration, hybrid sources of videos, unrealistic violent scenes, specificity of the context, etc., are the most remarkable commonalities. Moreover, all the presented datasets concern social violence such as street fights, movie fights or sport fights. Since there are no public campus datasets, few author designed campus violence experimental datasets [5–7], where the students were volunteers and experimental data were gathered by role-playing and daily life activities.

By analyzing methods and datasets from an extensive literature review, we understand that violence detection per se is an extremely difficult problem, since violence is a subjective concept in educational environments. Generally, violence detection techniques can fail due to actions which are wrongly interpreted as violent caused by fast movements and similarity with violent behaviors. The educational environment is the toughest challenge we faced because a machine learning algorithm is barely able to discern the difference between a violent hit or a hit for a joke. School and campus violence [7] differs from social violence since (1) the victim in campus violence events uses no weapons and (2) campus violence is generally not as strong as social violence. For this reason, campus violence can be sometimes confused with playing or sports with physical confrontation.

In line with existing and future methods, we also created a challenging dataset, specifically designed for shallow cost infrastructures that are mostly used nowadays in Italian public schools. Recent methods [24] may not be fully compatible with the infrastructure available in public schools, which could negatively affect the computational time. Changing illumination, complex backgrounds and low resolutions are a challenge for complex architectures as well [35]. In addition, the lack of data richness in datasets in the educational field is another important matter. Once we had analyzed the advanced action recognition techniques designed in the past, we built a lightweight 2DCNN system, trained with the proposed Daily School Break dataset. At a later stage, it was extended with ImageNet's datasets to increase the power and computational efficiency.

In the next Section, we present the proposed Daily School Break dataset in detail.

3. Daily School Break Dataset

To the best of our knowledge, currently there are no datasets containing video sequences acquired in the school yard without actor students. Moreover, the public availability of a similar dataset would provide support on different issues, including the detection of small and rapid violent actions.

By using a 14-megapixel Olympus Lens camera, we recorded 10 AVI/mp4 video files in the yard of a high school as listed in Table 2. To support video analysis, we provided in addiction five video files for testing. In total, the Daily School Break dataset contains 15 challenging videos for training and testing.

Table 2. Main characteristics of the videos contained in the DSB dataset.

Filename	Time (min:s)	Illumination Conditions	Type of Action
P7020001.avi	02:04	high	Violent (kick)
P7020002.avi	02:06	high	Violent (push)
P7020003-1.mp4	01:34	high	Non-violent
P7020003-2.mp4	01:45	high	Non-violent
P7020004-1.mp4	01.31	high	Non-violent
P7020004-2.mp4	01.37	high	Non-violent
P7020005-1.mp4	01:15	low	Non-violent
P7020005-1.mp4	01:18	low	Non-violent
P7020005-3.mp4	01:17	low	Non-violent
P7020006.avi	00:45	low	Non-violent

In detail, videos were acquired using three different wide angles: (a) seven videos aiming at the left corner of the school yard, (b) three videos pointing at the opposite side and (c) five recorded directly by high school students using a front-facing professional camera. The using of different wide angles allowed us to build a dataset containing heterogeneous video sequences. The videos were acquired on the same hour of the days (morning time) but with different wide angles, shadows and brightness. At first, the videos obtained from data acquisition stage were converted into frames with a frequency of 2 frame/sec. Afterwards, we decided not to apply any additional filters to adjust, smooth or refine data in order to replicate a realistic scenario. Thus, each sequence presents many challenging features which made the violence recognition a tough task.

In Figure 3, the main structure of the Daily School Break dataset is shown. All clip images are labeled as showing non-violent and violent behaviors, respectively. The images were organized into two main folders: (1) training set and (2) validation set. The images were sorted into their respective directories manually and stored randomly for creating training and validation datasets. Each folder was split into two other subfolders: (1) violent and (2) not violent. We used 65% of the total images for the training, while 35% were set for the validation phase. Specifically, 90% of the images are non-violent and include actions such as eating, chatting, walking, exulting and gesticulating (Figure 4a). In total, 10% of the scenes are rapid violent actions such as kicks, punches and hits (Figure 4b). In the final stage, we added new images for the test phase and binary classifications between violent and non-violent.

Collecting data regarding school environments faces many challenges due to school privacy and the lack of violent cases thanks to teacher surveillance, as well as organization and legal challenges. In our case, the overall cost of the built model should be more cost-effective thanks to the proper use of hardware implemented in the place where the SVD is applied. Therefore, the training model requires large and complex data, addictional time and computing resources which might not be available in public schools. Small and agile datasets can be used to detect school violence more effectively and freely than complex, costly and heavy datasets that require expensive operation, hardware and support. At the final stage, we incremented its capabilities with transfer learning techniques thanks to ImageNet models to increase its robustness.

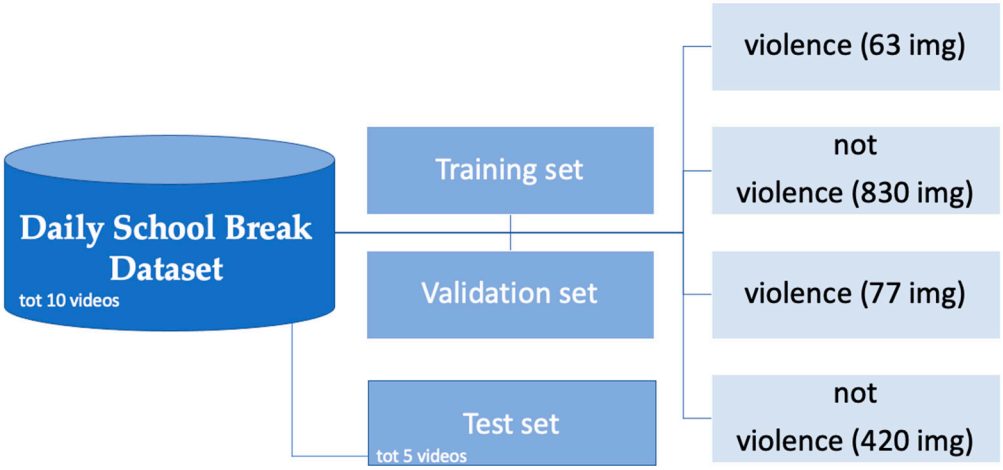
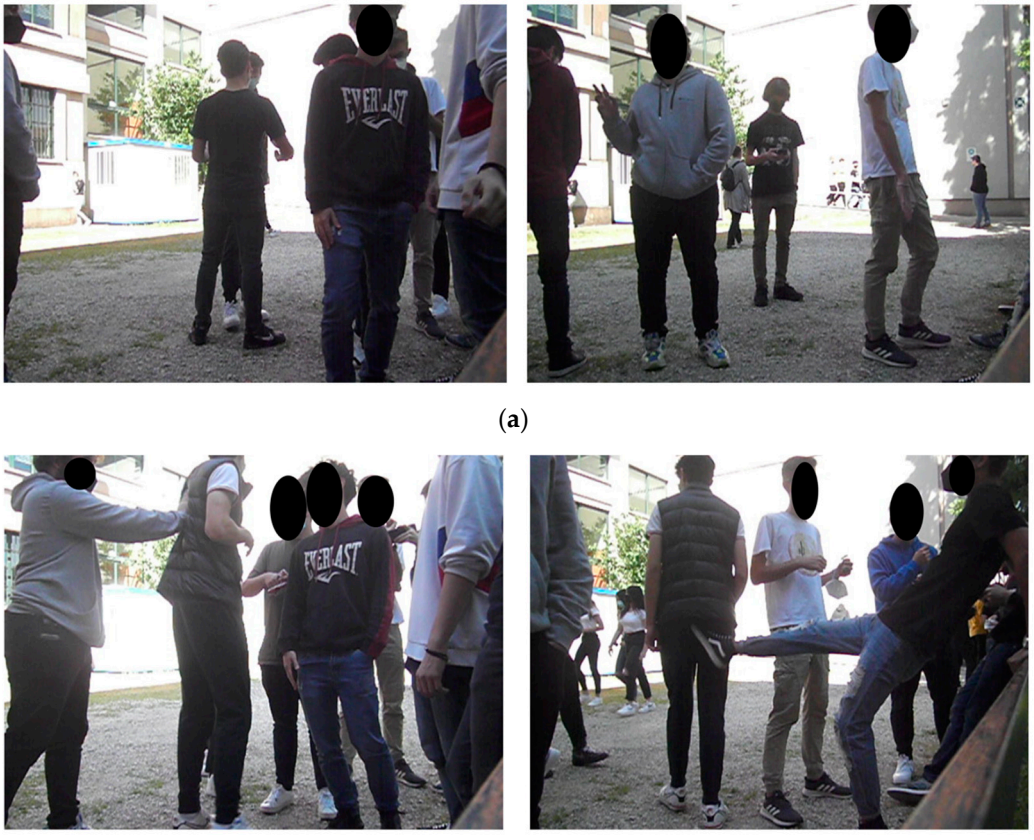


Figure 3. Logical architecture of the Daily School Break dataset.



(**a**)

(**b**)

Figure 4. (**a**,**b**) Daily School Break dataset: (**a**) non-violence and (**b**) violence sequence examples.

4. School Violence Detection system

In this Section, we provide a complete description of the School Violence Detection (SVD) system. From the hardware perspective, SVD is composed of an external Olympus camera for recording videos/images and an Intel core i5 2.5 GHz PC for processing.

A shallow infrastructure is used as the main device for training the deep learning system with 4 GB Ram. The hardware choice depends on the Italian public-school context, in which the available options are outdated with slow hardware and are affected by software failures. Obsolescence and no available updates can be serious problems, but the model is

supported by TensorFlow, which virtually trains models, datasets or parameters, while not affecting the ordinary PC performance too much.

From a logical perspective, the SVD system is based on a 2DCNN structure (Figure 5). The software components of SVD are based on the piloting API interface Keras as the frontend and TensorFlow for the access to the backend.

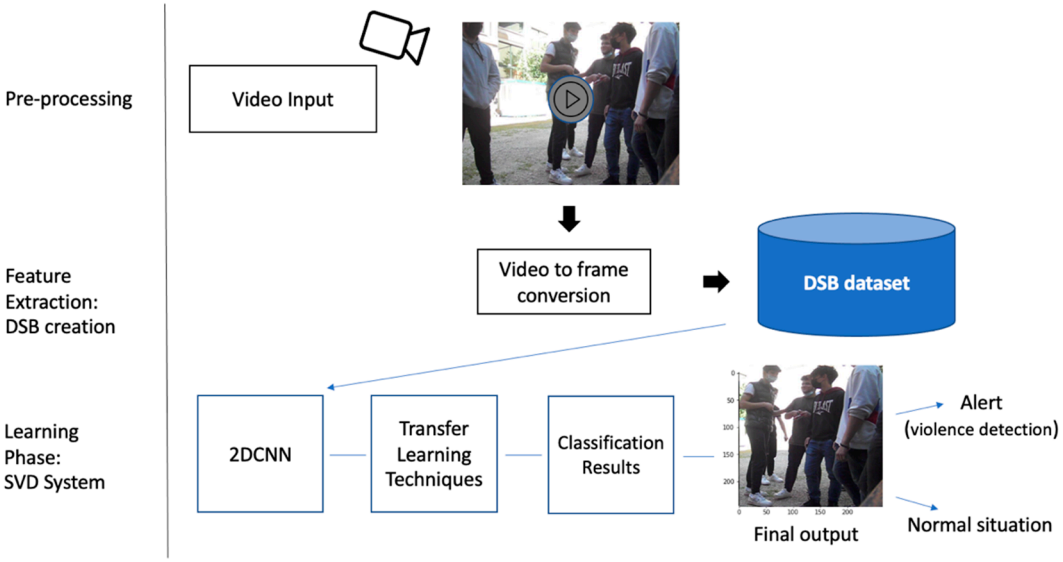


Figure 5. Logical architecture of School Violence Detection (SVD) system.

The goal of SVD is to create an easy, computationally efficient ML model able to classify violent and non-violent images within school setting.

We trained our system with the Daily School Break dataset and in a later stage we enriched the SVD with transfer learning techniques on the ImageNet model (VGG19). Lower power consumption and higher throughput training are difficult to implement on lightweight hardware without any external model. The result is an useful prototype as a baseline reference for future works that can be compared by using the proposed dataset and in other applications on educational environments.

In detail, each DSB input image was resized to 224 × 224 pixels. No other pre-processing procedure was applied. After resizing the images, a sequential model architecture was constructed. The model first takes the pre-processed image with the right shape. The created feature map is the input of the convolutional layer: we implemented 32 filters at the first stage plus ReLu activation function to prevent negative pixel values. In the following layers, the number of filters increased: 64, 128 and up to 512 neurons. Then, we added a max-pooling layer to reduce the features map, leaving the important features to be detected by the model. It converts pooled feature maps from 2D arrays into 1D, preparing them to be used by the fully connected layer. At a later stage, we implemented a fully connected output layer (Dense Layer) with a sigmoid activation function to normalize output values between zero and one. This function is required since the prediction outcome is binary: violent (Figure 6a,b) and non-violent (Figure 6c,d). The output of each neuron at this stage represents the probability of the sample belonging to each respective class. In the end, to evaluate the model, in the run phase we also considered three parameters: (1) an optimization function (in this case the "adam" algorithm), (2) a loss function (a "binary crossentropy" for the binary outcome classifier (Figure 5)) and (3) the accuracy as a performance metric.

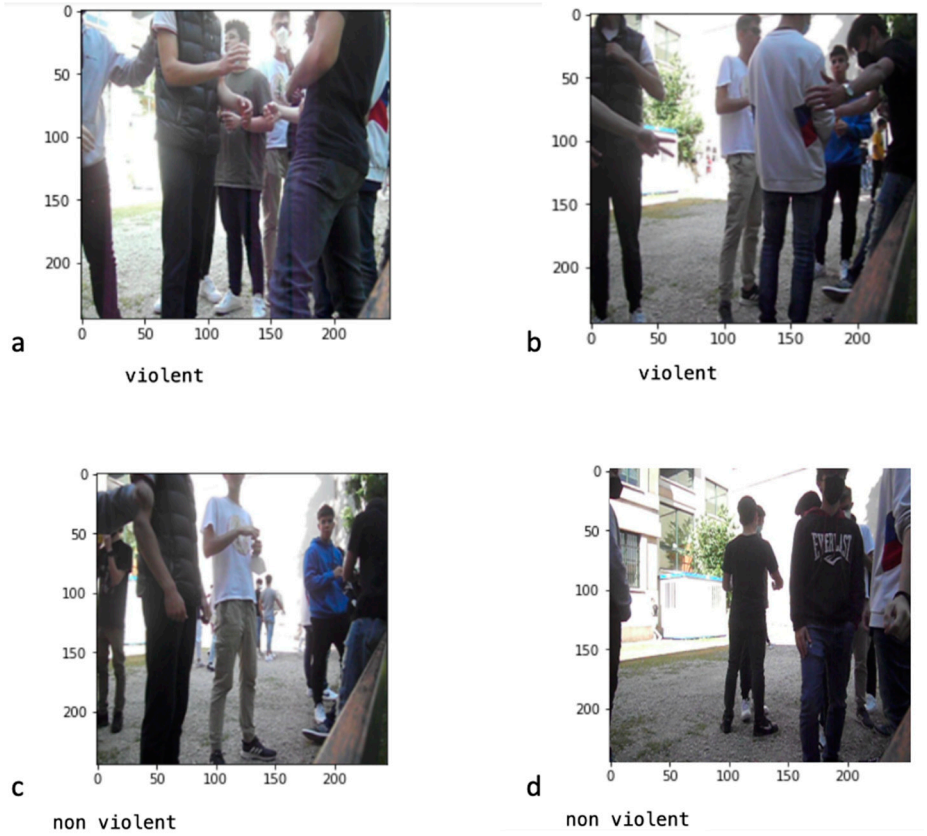


Figure 6. Running examples of the final output of the SVD that recognizes (**a**,**b**) violent and (**c**,**d**) non-violent scenes.

Concerning SVD performance, as shown in Figure 7a, we noticed a sharp change in accuracy level during the training phase: performance can drop dramatically due to over-fitting that affects small datasets. Thus, we applied transfer learning techniques to make the model stronger in classification tasks. Briefly, transfer learning allows us to analyze a new image by exploiting the properties of the previously acquired knowledge to recognize further objects or actions. The learning process no longer starts from scratch, as the weights are reused to solve a new activity easily. The main advantages of applying transfer learning are saving time in training, expectation of better performance and error reduction.

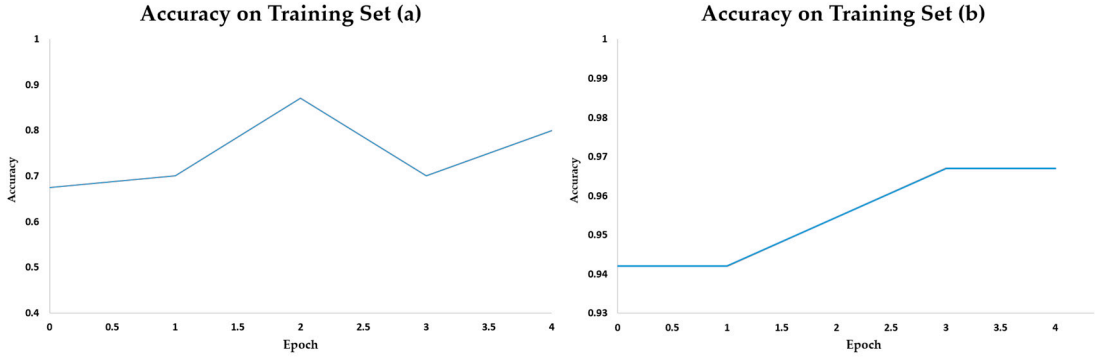


Figure 7. Level of accuracy during the training phase obtained on (**a**,**b**) DSB datasets.

As shown in Figure 7, an epoch is the number of passes a training dataset takes to learn and adjust the internal model parameters. Integrating transfer learning to SVD, as shown in Figure 7b, gives more accuracy and stability to our prototype during the training phase.

Indeed, transfer learning is so powerful because it uses the weight from the pre-trained model to initialize the weight of the prototype, joining them in a new and more precise prototype. Therefore, the using of a pre-trained model saves precious time; models trained on ImageNet perform well in real-world image classification problems because they contain over one million samples. Thus, in the extended SVD prototype, every epoch requests only 85s (3 s/step) instead of 240s (9 s/step) without the transfer learning process.

The proposed extended SVD is the result of the previous evaluation of three pre-trained ImageNet networks (VGG16 [9], VGG19 and ResNet50). We tested the three models in parallel. With the first epoch, applying VGG16, we achieved an accuracy of over 94%. Without changing any parameters or iterators, we obtained an accuracy of more than 95% with VGG19 and around 96% with ResNet-50. In all these models, the accuracy stabilizes itself towards the latest processing. Unfortunately, the transfer learning technique exhibits limitations, and the overfitting problem persists. In addition, applying ResNet-50 transfer learning, the output results were not acceptable due to the high number of false positives. ResNet-50 transfer learning was the worst one because the validation loss increased with the validation accuracy. In detail, the transfer learning process did not give a suitable solution. ResNet-50 was too powerful for our DSB dataset. Later, we further evaluated SVD extended with ResNet-50 in order to log the score for image prediction. The experiment is shown in Figure 8. Images with a score lower than 0.5 represent non-violent images, while the others (with a score greater than 0.5) represent violent images. Although we achieved satisfactory performance results, there are still many non-violent images that are mistakenly classified as violent.

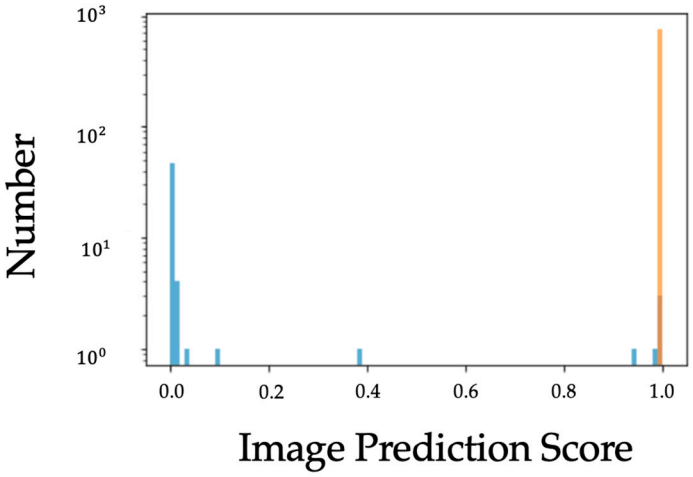


Figure 8. Histogram plot shows predicted DSB non-violent (light blue bars) and violent (orange bars) actions in images classified in the prototype trained with SVD enriched with ResNet-50.

We tried to remove some layers without derationing the output as demonstrated in [32], but we obtained low results. In addition, we discovered that the features learned from the higher levels are not sufficient to differentiate the classes of the proposed problem, e.g., the trained model identified close people, but it was not able to understand a push from a pat on the back.

As visualized in Figure 9, by applying VGG19 we obtained the best performance: we saw that the validation loss decreased and the validation accuracy increased with every epoch. This means that the model was learning correctly. Eventually, we chose VGG19 that performs better with our DSB dataset reaching an accuracy of 95%. In detail, the process requested several steps: (1) pre-process the image as the correct input for the pre-trained network, (2) remove the classification layer, (3) freeze the weights of the network, (4) add a new classification layer combining the proposed 2DCNN and the pre-trained model (VGG19) into one and (5) train again the extended SVD with the DSB dataset.

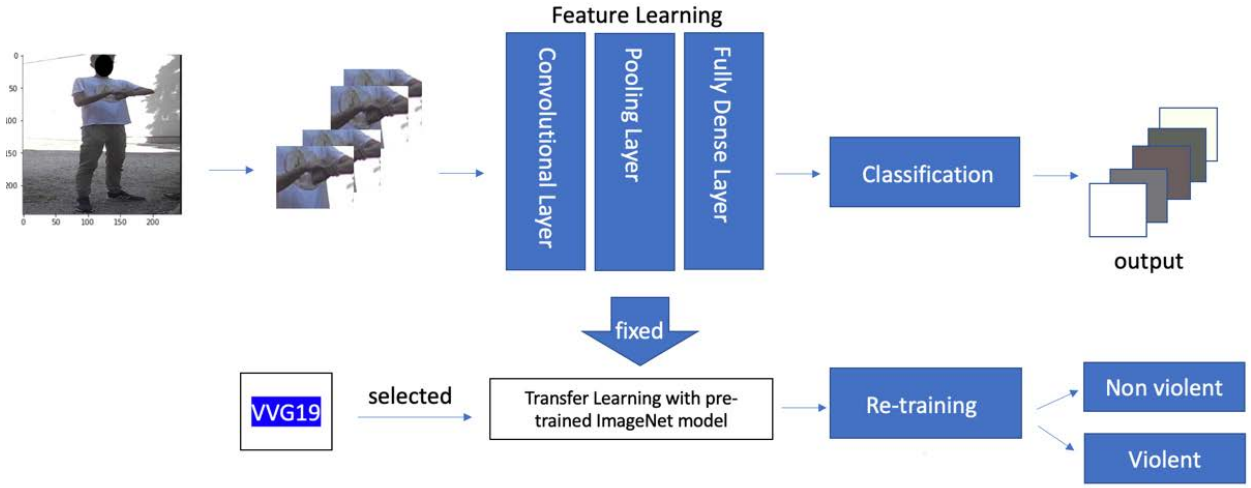


Figure 9. Logical architecture of the extended SVD enriched with transfer learning with pre-trained ImageNet model.

5. Experimental Results

This Section describes the results of the proposed SVD applied to DSB dataset. Regarding the performance evaluation, we measured the number of frames in which the classification is not properly detected.

Loss and accuracy are two metrics that are useful for counting the number of frames in which true/false positives (TP, FP) and true/false negatives (TN, FN) [36] are recognized. In detail, loss can be seen as the distance between the true values of the problem and the values predicted by the model. In our model, the probability to predict violent or non-violent classes as binary ranges between one and zero. For example, if the probability of a test violent scene is 0.6, the probability of non-violent is 0.4. In this case, the image is classified as violent. Briefly, loss will be the sum of the difference between the predicted probability of the real class of the test picture and one. Accuracy (A) can be seen as the percentage of data classified correctly. The equation for A is as follows [36]:

$$A = \frac{TP + TN}{TP + TN + FP + FN}$$

Table 3 shows some visual metrics (A and loss) comparing original SVD with extended SVD.

Table 3. A and loss on original and extended SVD.

	A (%)	**Loss (%)**
Original SVD	85	48
Extended SVD	95	26

Experimental results in [37] over the datasets described in Section 2.2 witnessed stable performance in violence detection tasks. We tried to demonstrate that the proposed SVD prototype can achieve balanced results for the DSB dataset too. For each presented SVD system (original (Figure 10a) and extended system (Figure 10b)), the performances are pointed out by plotting the confusion matrix in the form of a heat map. Figure 10 summarizes the prediction results in percentage testing random DSB images. We found that original DSB images are mostly outdoors, and that they are covered by lighting issues such as shadows or blurring which affect accuracy. Moreover, transfer learning techniques increase the precision of the SVD system helped by the adding of new images via web scraping. Indeed, online indoor scenes with students often have fewer lighting issues. A static camera and no obstacles can ensure that there are not any background subtractions; in this way, it is easier to detect violent scenes.

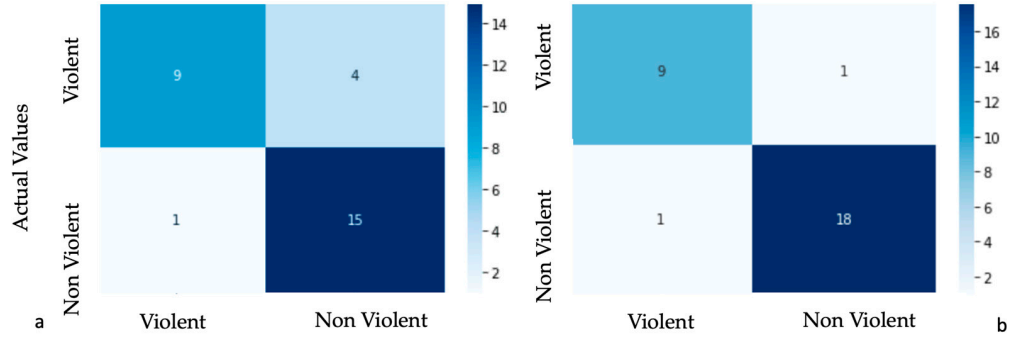


Figure 10. Visualization of the confusion matrix of the original (**a**) and extended (**b**) SVD.

We gained good results with the last model (extended SVD), recognizing violent or non-violent actions in less than two minutes with an A of 95% using shallow cost hardware (Lenovo PC intel core i5).

In addition, we needed to conduct more extensive testing on our prototype to analyze the performance on different subjects and objects. To evaluate our extended SVD with a bigger dataset, we decided to combine our DSB with new images downloaded via the web scraping mechanism. We used specific keywords such as "fighting kids", "bullying students", "violence kids", etc., surfing on Google Image to implement DBS with violent images regarding school environments (Figure 11). As previously mentioned, the images were manually analyzed and split into the correct directories: non-violent and violent. To sum up, the combined dataset was made of 1800 images: 83% of them came from original samples collected in educational settings and the remaining 17% via web scraping. The extended SVD was trained with the combined dataset, and it achieved an A of 95% and loss of 26.

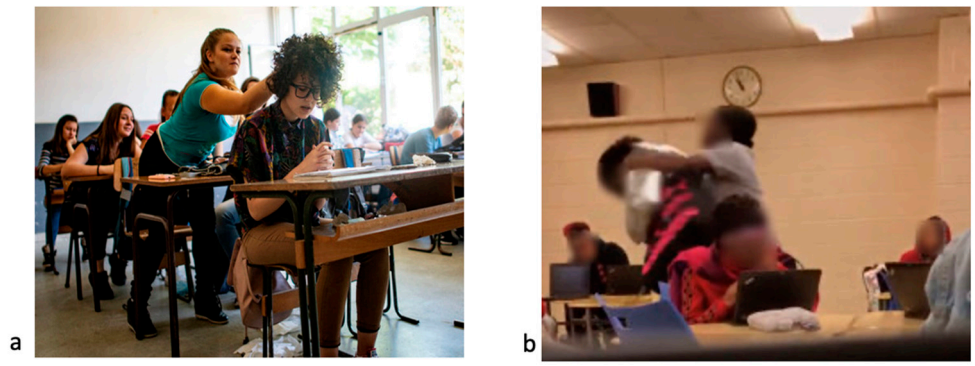


Figure 11. Samples of "violence student bullying" on (**a**) and "violence student fight" on (**b**).

As a young field, SVD system is currently under the evaluation phase in order to be implemented and used in the schools. In public environments, violent detections in educational contexts can represent a possible severe threat to the social stability that should be deeply investigated.

A comparative analysis about DSB image quality is presented in the next section.

Robustness Evaluation of DSB Images

An additional experimental session has been carried out with the purpose to assess the model analysis of the original DSB images. In particular, we used Captum, a Pytorch tool and ViT (Figure 12) to research the classifications computed by the SVD system. We analyzed how the original SVD recognized the actions. By applying algorithms such as gradient, semantic segmentation [18] (Figure 13) and occlusion analysis [30] (Figure 14), we tried to indagate which areas of the image contributed to generate the final neural model outputs.

Figure 12. Original DSB image analysis using ViTFeatureExtractor. The class recognized by ViT is "bars": the model pays attention to the foreground without considering the violent scene (push) in the center part of the image. The difficulty of image processing mainly depends on the nature of the videos or on the speed of the action in the scene, as in punches or hits, which in the reality are fast and uncontrolled. Although different extended techniques and mechanisms have been applied, the just-mentioned limitations highlight the problem of analyzing a real environment such as a school yard.

Figure 13. Original DSB image analysis using sematic segmentation.

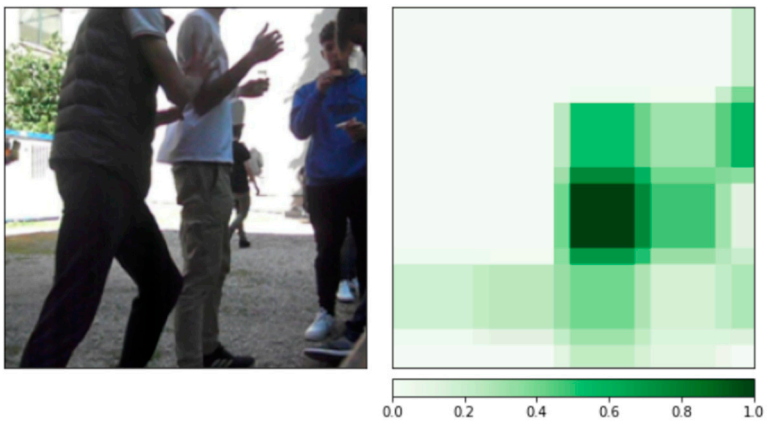


Figure 14. Original DSB image analysis using occlusion technique in which the model tries to classify by coloring in green the critical areas for the decision.

As shown in Figure 13, semantic segmentation analysis colors the pixels that contribute to label a particular class. It identifies regions that include humans in the front, while neglecting objects and people in the background, creating confusion in the original image.

The problem of understanding the aspect of visual appearance has become particularly relevant. In [32], the growing interest in feature visualization and attribution is shown. The saliency map is applied on Figure 15 in order to know how a neural network interprets for future improvements and corrections.

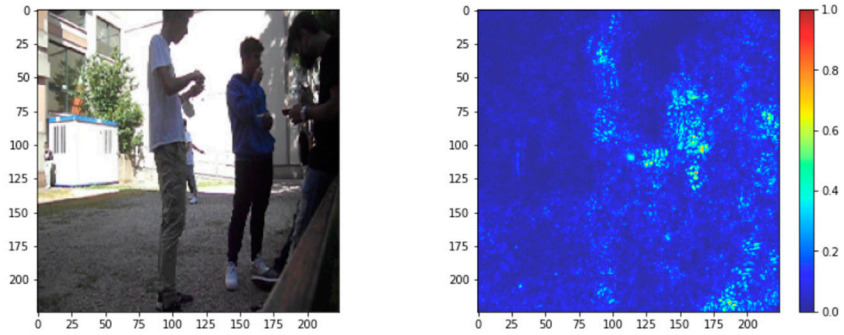


Figure 15. Original DSB image analysis using attribution, also known as a saliency map [38]. It is a technique to rank the pixels that highlight the areas of the given image, discriminating the zones using colors. Cyan shows the areas used to compute: the student in the middle and the other two classmates on the right contribute more to detection.

6. Conclusions

In this paper, a 2DCNN-based model is applied to design and develop a SVD system able to detect (violent or non-violent) actions in schools. The system is implemented on shallow cost architectures. We extend SVD capabilities using VVG19, a pre-trained neural network. A Daily School Break dataset is presented as an innovation in the violence detections dataset context where actual students without any filter are presented. It consists of 15 videos acquired in a school yard with such different characteristics as lighting, shapes and people. We tried to fill a not yet covered gap in the analyzed papers [3,5,12] with a special attention on low-cost solutions. We proposed a dataset collected in real-life public schools, focusing on reduced computational load and improving usability directly in schools. Our dataset is designed to be light and not computationally expensive which is beneficial to use in time-sensitive applications or in edge devices. Experimental results show that by refining the dataset via web-scraping and applying transfer learning techniques, the classification of the network achieves an A around 95%. Some obstacles still remain, i.e., the noise in the images and the difficulty to obtain images with minors as the main subjects. The paper also provides different metrics to compare the A of the results obtained in shallow cost architectures.

In the future, we could enhance our SVD by analyzing the time series of frames or implementing original audios using the DSB dataset. Furthermore, this model could be integrated on interdisciplinary fields such as psychology and education to create a complete analysis.

Author Contributions: Conceptualitazion, methodology, software, validation, E.P. and G.L.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The original Daily School Break dataset can be downloaded from this link (https://uniudamce-my.sharepoint.com/personal/123767_spes_uniud_it/_layouts/15/onedrive.aspx?id=%2Fpersonal%2F123767%5Fspes%5Funiud%5Fit%2FDocuments%2FDATASET%2DSVD%2DPAPER&ga=1 (accessed on 22 February 2023)).

Acknowledgments: This work was partially supported by the Departmental Strategic Plan (PSD) of the University of Udine—Interdepartmental Project on Artificial Intelligence (2020-25).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sudhakaran, S.; Lanz, O. Learning to detect violent videos using convolutional long short-term memory. In Proceedings of the 14th IEEE International Conference on Advance Video and Signal Based Suirveillance (AVSS), Lecce, Italy, 29 August–1 September 2017; pp. 1–6.

2. Accattoli, S.; Sernani, P.; Falcionelli, N.; Mekuria, D.N.; Dragoni, A.F. Violence Detection in Videos by Combining 3D Convolutional Neural Networks and Support Vector Machines. *Appl. Artif. Intell.* **2020**, *34*, 329–344. [CrossRef]
3. Cheng, M.; Cai, K.; Li, M. RWT-2000: An open large scale video database for violence detection. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milano, Italy, 10–15 January 2021.
4. Nievas, E.B.; Suarez, O.D.; Garcia, G.B.; Sukthankar, R. Hockey fight detection dataset. In *Computer Analysis of Images and Patterns*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 332–339.
5. Bianculli, M.; Falcionelli, N.; Sernani, P.; Tomassini, S.; Contardo, P.; Lombardi, M.; Dragoni, A.F. A dataset for automatic violence detection in videos. *Data Brief* **2020**, *33*, 106587. [CrossRef]
6. Xing, Y.; Dai, Y.; Hirota, K.; Jia, A. Skeleton-based method for recognizing the campus violence. In Proceedings of the 9th International Symposium on Computational Intelligence and Industrial Applications, Beijing, China, 19–20 December 2020.
7. Ye, L.; Liu, T.; Han, T.; Ferdinando, H.; Seppänen, T.; Alasaarela, E. Campus Violence Detection Based on Artificial Intelligent Interpretation of Surveillance Video Sequences. *Remote. Sens.* **2021**, *13*, 628. [CrossRef]
8. Calzavara, I. *Human Pose Augmentation for Facilitating Violence Detection in Videos: A Combination of the Deep Learning Methods DensePose and VioNet*; Department of Information Technology and Media (ITM), Mid Sweden University: Sundsvall, Sweden, 2020.
9. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
10. Xiao, J.; Wang, J.; Cao, S.; Li, B. Application of a Novel and Improved VGG-19 Network in the Detection of Workers Wearing Masks. *J. Phys. Conf. Ser.* **2020**, *1518*, 012041. Available online: https://iopscience.iop.org/article/10.1088/1742-6596/1518/1/012041 (accessed on 9 November 2022). [CrossRef]
11. Sumon, S.A.; Goni, R.; Bin Hashem, N.; Shahria, T.; Rahman, R.M. Violence Detection by Pretrained Modules with Different Deep Learning Approaches. *Vietnam. J. Comput. Sci.* **2019**, *7*, 19–40. [CrossRef]
12. Sultani, W.; Chen, C.; Shad, M. Real-word anomaly detection in surveillance videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6479–6488.
13. Bermejo, E.; Deniz, O.; Buono, G.; Sukthankar, R. Violence Detection in Video Using computer Vision Techniques. In Proceedings of the International Conference on Computer Analysis of Images and Patterns, CAIP 2011, Seville, Spain, 29–31 August 2011; pp. 332–339.
14. Yun, K.; Honorio, J.; Chattopadhyay, D.; Berg, T.L.; Samaras, E. Two person interaction detection using body pose features and multiple distance learning. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–22 June 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 28–35.
15. Perez, M.; Kot, A.C.; Rocha, A. *Detection of a Real Word Fights in Surveillance Videos*; IEEE: New York, NY, USA, 2019; pp. 2662–2666.
16. Vijeikis, R.; Raudonis, V.; Dervinis, G. Efficient Violence Detection in Surveillance. *Sensors* **2022**, *22*, 2216. [CrossRef]
17. Choqueluque-Roman, D.; Camara-Chavez, G. Weakly Supervised Violence Detection in Surveillance Video. *Sensors* **2022**, *22*, 4502. [CrossRef]
18. Dong, Z.; Qin, J.; Wang, Y. *Multi-Stream Deep Networks for Person to Person Violence Detection in Videos*; Tan, T., Li, X., Chen, X., Zhou, J., Yang, J., Cheng, H., Eds.; Pattern Recognition. CCPR 2016. Communications in Computer and Information Science; Springer: Singapore, 2016; Volume 662. [CrossRef]
19. Demarty, C.-H.; Penet, C.; Soleymani, M.; Gravier, G. VSD, a public dataset for the detection of violent scenes in movies: Design, annotation, analysis and evaluation. *Multimed. Tools Appl.* **2014**, *74*, 7379–7404. [CrossRef]
20. Dandage, V.; Gautam, H.; Ghavale, A.; Mahore, R.; Sonewar, P.A. Review of Violence Detection System using Deep Learning. *Int. Res. J. Eng. Technol.* **2019**, *6*, 1899–1902.
21. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, Austria, 3–7 May 2021.
22. Fraga, S. Methodological and ethical challenges in violence research. *Porto Biomed. J.* **2016**, *1*, 77–80. [CrossRef]
23. Ramirez, H.; Velastin, S.A.; Meza, I.; Fabregas, E.; Makris, D.; Farias, G. Fall Detection and Activity Recognition Using Human Skeleton Features. *IEEE Access* **2021**, *9*, 33532–33542. [CrossRef]
24. Ullah, F.U.M.; Obaidat, M.S.; Ullah, A.; Muhammad, K.; Hijji, M.; Baik, S.W. A Comprehensive Review on Vision-Based Violence Detection in Surveillance Videos. *ACM Comput. Surv.* **2023**, *55*, 1–44. [CrossRef]
25. Wang, W.; Dong, S.; Zou, K.; Li, W. A Lightweight Network for Violence Detection. In Proceedings of the 2022 the 5th International Conference on Image and Graphics Processing (ICIGP 2022), Beijing, China, 7–9 January 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 15–21. [CrossRef]
26. Ullah, F.U.M.; Obaidat, M.S.; Muhammad, K.; Ullah, A.; Baik, S.W.; Cuzzolin, F.; Rodrigues, J.J.P.C.; de Albuquerque, V.H.C. An intelligent system for complex violence pattern analysis and detection. *Int. J. Intell. Syst.* **2022**, *37*, 10400–10422. [CrossRef]
27. Su, Y.; Lin, G.; Zhu, J.; Wu, Q. Human interaction learning on 3d skeleton point clouds for video violence recognition. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 74–90.
28. De Boissiere, A.M.; Noumeir, R. Infrared and 3d skeleton feature fusion for rgb-d action recognition. *IEEE Access* **2020**, *8*, 168297–168308. [CrossRef]
29. Deng, J.; Dong, W.; Socher, R.; Fei-Fei, L. ImageNet: A large scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 255–258.
30. Talo, M. Convolutional Neural Networks for Multi-class Histopathology Image Classification. *arXiv* **2019**, arXiv:1903.10035.

31. Veit, A.; Wilber, M.; Belongie, S. Residual networks behave like ensembles of relatively shallow networks. In Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16), Barcelona, Spain, 5–10 December 2016; Curran Associates Inc.: Red Hook, NY, USA, 2016; pp. 550–558.
32. Olah, C.; Mordvintsev, A.; Schubert, L. Feature Visualization. How neural networks build up their understating of images. *Distill* **2017**, *2*, 0007. Available online: https://distill.pub/2017/feature-visualization (accessed on 28 March 2023).
33. Hassner, T.; Pitcher, Y.; Kliper-Gross, O. Violent flows: Real time detection of violent crowd behavior. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 1–6.
34. Ullah, F.U.M.; Ullah, A.; Muhammad, K.; Haq, I.U.; Baik, S.W. Violence Detection Using Spatiotemporal Features with 3D Convolutional Neural Network. *Sensors* **2019**, *19*, 2472. [CrossRef]
35. Varga, D. No-Reference Image Quality Assessment with Convolutional Neural Networks and Decision Fusion. *Appl. Sci.* **2021**, *12*, 101. [CrossRef]
36. Avola, D.; Cinque, L.; Foresti, G.L.; Martinel, N.; Pannone, D.; Piciarelli, C. A UAV Video Dataset for Mosaicking and Change Detection from Low-Altitude Flights. *IEEE Trans. Syst. Man Cybern. Syst.* **2018**, *50*, 2139–2149. [CrossRef]
37. Mumtaz, N.; Ejaz, N.; Aladhadh, S.; Habib, S.; Lee, M.Y. Deep Multi-Scale Features Fusion for Effective Violence Detection and Control Charts Visualization. *Sensors* **2022**, *22*, 9383. [CrossRef] [PubMed]
38. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep Inside Convolutional Networks: Visualizing Image Classification Models and Saliency Maps. *arXiv* **2013**. [CrossRef]