Can We Ensure Safety? Violence Detection in Video Scenes

Flavien Vidal
École Polytechnique
flavien.vidal@polytechnique.edu

Disclaimer—This work is a four-hour in-class data competition and, for that reason, does not claim to be a complete solution to detecting violence in videos.

***Abstract*—The amount of visual content published online, especially through social media, has dramatically increased in recent years. As a result, the risk of users being exposed to discriminatory, sexual or violent content has also increased. It is therefore becoming increasingly important to develop effective and flexible approaches to detect content that violates the usage policies of such networks. In this short project, we focus on the detection of violent content. In particular, we are only interested in visual contents although audio and textual contents provide valuable additional information. We use a public dataset, namely the Real Life Violence Situations Dataset (RLVS), to train classical learning models such as SVM and XGBoost as well as deep neural architectures including hybrid CNN-SVM approaches and MobileNet and Inception-based networks. The experimental results demonstrate the effectiveness of the different proposed models by achieving precision and recall up to 94.97% and 96.17% respectively on this dataset.**

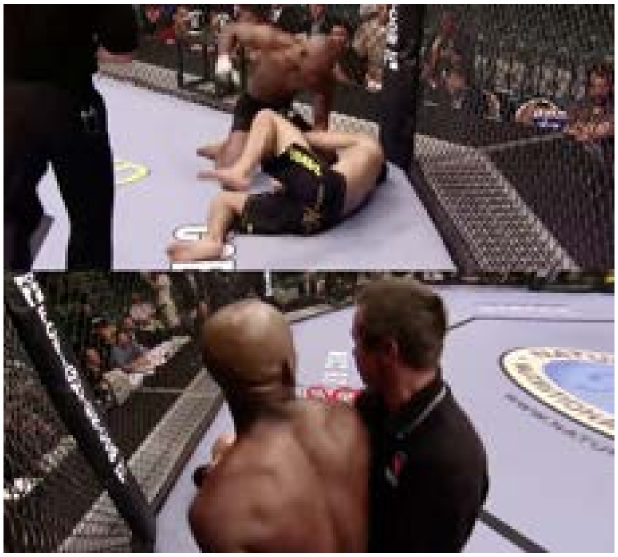


Figure 1: Our method estimates, from the frames of a video clip, if a content is potentially harmful or not. This information can then be used to better fight against their behaviors and reduce the risk of collision with a vigilante. On this specific example, the upper frame is correctly classified as violent by our model while the lower one is classified as safe.

I. Introduction

In recent years, video data has become ubiquitous and this is partly due to the growth of many online platforms where people upload tons of data, as well as due to the surveillance cameras deployed around the whole world that also collect a huge amount of data. For example, these cameras were generating about 566 petabytes of video data per day in 2016 and YouTube users were uploading almost 300 hours of video per minute. As a consequence, a major challenge lies in monitoring this massive amount of video data. In particular, a system capable of automatically detecting violence in both video games, Hollywood movies or real life recordings becomes extremely useful. The failure to detect these violent contents can sometimes be highly prejudicial and, unfortunately, it can prove to be a very complicated task in some circumstances. For example, the inherent biases of datasets have a major influence on the models' predictions and can lead to sexist or racist behavior. They may also prevent models from recognizing less common but extremely serious forms of violence, such as killing and war scenes that do not necessarily contain physical combat. The difficulty also lies in the fact that many other apparently violent videos are actually legal, for example because they are taken from video games or from Hollywood movies. It is therefore necessary to be able to detect violence according to the nature of these videos.

In this short project, we propose a simple but efficient solution to detect violent content within video clips. In particular, we intend to classify frames within videos as violent or safe. The study of the sequential aspect of videos is not to be addressed in this project. The study of the sequential aspect of videos, which however accounts for an essential part of violence recognition in videos, is not to be addressed in this competition. In this case, only visual features are used to detect violence, we do not care about audio features which should however bring additional valuable information.

SVM, XGBoost, standard CNN, hybrid CNN-SVM architectures as well as MobileNet- and Inception-based models are trained on these visual features to detect violence. A publicly available dataset, namely the Real Life Violence Situations Dataset (RLVS), is used for classifier training and testing. Results achieved with the Inception-based model are better than the results of the other tested models and are very promising. Moreover, it exhibits strong generalization properties.

Section II. motivates the choice of the dataset. In particular, we make sure it enables to produce generalizable models, without bias and especially discriminatory bias, as we believe this is an extremely important factor. Section III. discusses the different methods frequently used for violence detection in videos. Since time is limited we heavily rely on these methods to ensure that we achieve decent results quickly. In addition we describe the different data augmentation techniques we have used. Finally section IV. describes the experiments in more detail. We present the experimental setup, the learning process, the metrics on which we rely and base our conclusions and a comparison of the results obtained.

II. Real Life Violence Situations Dataset

As mentioned previously, in this project we are using the Real Life Violence Situations Dataset[1] that was created due to the lack of publicly available datasets related to violence between individuals [**?**]. It contains 1000 violent and 1000 non-violent RGB videos that were collected from YouTube. The violent videos are collected from many real situations of street fights in different environments and conditions. The non-violent videos are collected from many different human actions, such as sports, eating, walking, etc.

A. Data exploration

In order to have more insight into the nature of the given videos, we decided to create a mosaic consisting of a single frame of n different randomly drawn videos (with n ranging from 200 to 800). Since this project was limited in time, this mosaic saved us a precious amount of time because we did not have to watch dozens of videos to get an idea of the overall quality and diversity of the dataset. *Figure* 2 shows an example of one of these mosaics. A more representative example based on 200 random violent videos is given in Appendix.

After investigating these different frames, we found that the quality varied greatly depending on the videos but also that the data were fairly diverse in nature, which is a crucial factor. For example, within the violent records, we find both street fights and professional fights, and ranging from two to a dozen opponents. Videos presented in *Figure* 3 illustrate some of these different natures of violence.

However, it seems to us that the large majority of these films only depict physical violence such as fights and brawls, but no or few more serious violence. As mentioned in the introduction, our model will therefore potentially be unable to detect scenes of extreme violence such as murders or wars. This can lead to major issues. Similarly, it seems to us that the dataset includes little or no cartoon violence. Our model may therefore not be generalizable to this kind of violence. To some extent, augmentation techniques that convert the style of images into a more cartoony style could partially address this problem.

[1] Dataset is available at `https://paperswithcode.com/dataset/real-life-violence-situations-dataset`

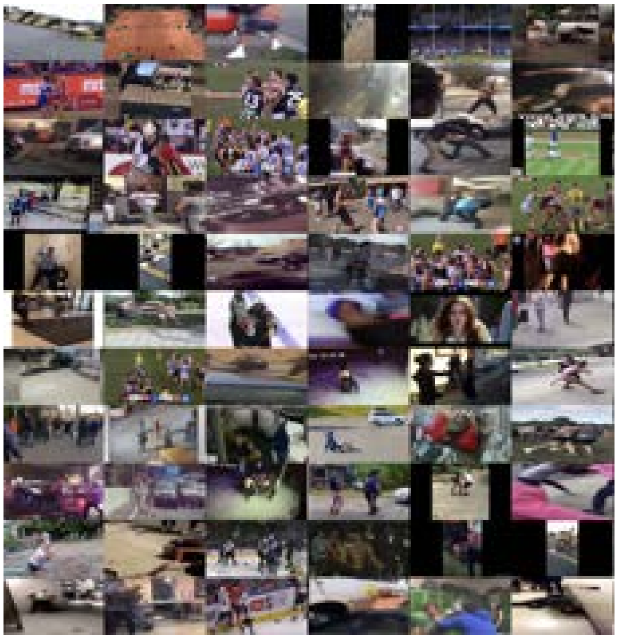


Figure 2: Subset of a mosaic made of single frames randomly taken from 66 different videos.

Regarding non-violent videos, we found that a large majority came from sports programs. Few of them come from everyday life which can also be problematic. Appendix shows an example of situation from the daily life as well as more frequent examples from high jump, weight lifting and golf.

B. The Risk of Bias in Violence Detection

Violence detection systems can be used in many sensitive environments to make important decisions. It is therefore critical to ensure that these decisions do not reflect discriminatory behavior toward certain groups or populations.

Indeed, algorithms can be vulnerable to bias and make unfair decisions by favoring an individual or a group of individuals. A famous example is the use by U.S. courts of a software program that measures a person's risk of re-offending. An investigation found that the software was more likely to have higher false-positive rates for African-American convicts than for Caucasians, falsely predicting that they had a higher risk of recidivism. Similar flaws have been found in many other areas. It is therefore important to keep in mind the importance of equity through our project.

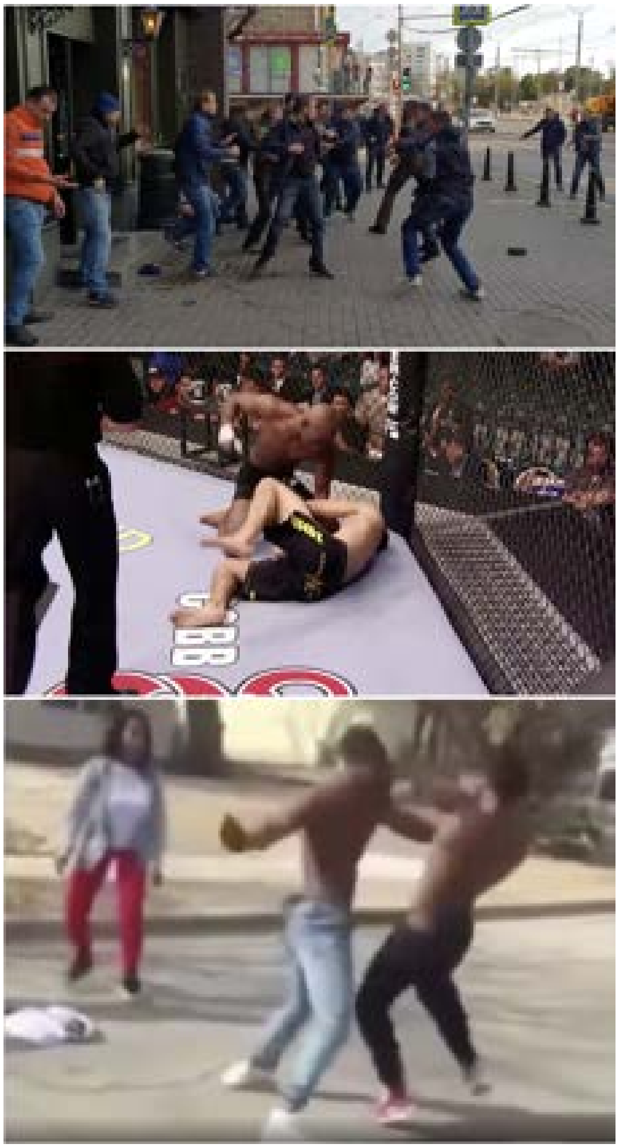


Figure 3: Highlight of different kind of violence in videos V407, V760 and V855 (from top to bottom).

There are two possible sources for such unfairness, those arising from data and those arising from algorithms. Therefore, we started by identifying the biases the dataset was likely to induce in our predictions. We observed that both violent and non-violent videos included fights from different ethnic communities, so that hopefully our algorithm will not be biased against certain ethnicities. We also found that most of the violent videos appeared to be of men.

Nevertheless, we believe that this will not affect the learning process since these variations in frames are only very local.

C. Additional statistics

Here are some additional statistics that we computed and that subsequently guided some of our modeling choices.

property	**frame count**		**frame rate**	
class	violent	safe	violent	safe
mean	159.78	127.59	29.56	24.87
std	374.08	168.90	3.55	5.13
min	62	29	10.50	11.00
med	147	125	29.97	25.00
max	11272	5397	37.00	30.02

Table 1: Statistics on the total number of frames and frame rate per nature of video.

III. Violence Detection

In this section, we first describe the methods commonly used for the task of detecting violence in videos and then detail the approach we used, in particular the data augmentation techniques and the models.

A. Violence detection techniques

Figure 4 shows the percentage of use of different methods for the task of detecting violence in videos. We observe that SVM is a common method in this field and accounts for almost a quarter of all methods. Conventional methods used from 2015 to 2018 account for about one fifth of all approaches and forest-based methods account for a little more than 10% of them. Since the increase in the performance of computing equipment, deep learning techniques are becoming more common in violence detection. Overall, 43% of all applied methods use deep learning, the vast majority of which are based on convolutional neural networks.

B. Data augmentation

To make reliable predictions, deep learning models require a lot of training data. The available dataset has 2000 videos evenly distributed across the two classes. We chose to augment these existing data to not only increase the size of the set to create a better model but also to avoid overfitting. By increasing the size of the data and adding diversity, we hope to help the model generalize better hence preventing overfitting.

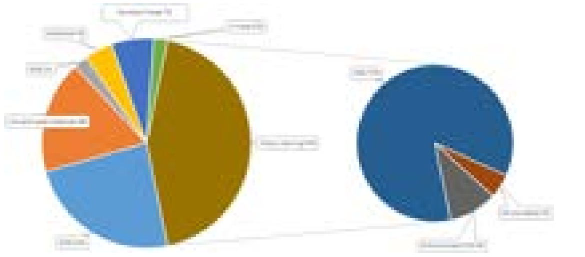


Figure 4: Distribution of commonly used techniques for detecting violence [?].

Among the existing augmentation techniques, we chose to experiment a dozen of them. We started by combining smooth augmentation methods, then more severe ones and compared the results obtained from the same models. Some of the techniques we experimented with on the input frames consisted in inverting them horizontally with a probability value ranging from 0.5 to 1.0, resizing them between 75 and 125% of their original size, doing the same but independently for each axis, moving them from -20 to $+20\%$ on the x and y axes independently, rotating them by $-25°$ to $+25°$, shearing them by $-16°$ to $+16°$ degrees, multiplying the brightness of 90% of them by a random value between 0.5 and 1.5 and multiplying the remaining 10% channel-wise.

Additionally, we applied distortions and in particular affine transformations that differ between local neighborhoods (PiecewiseAffine) [?]. For this, a regular grid of points is set up on the considered image and the neighborhood of these points is randomly moved by means of affine transformations. This leads to local distortions. We also transformed frames by locally moving pixels using displacement fields (ElasticTransformation). Studies have shown that a 10:1 ratio of hyperparameters α and σ seems to be satisfactory, e.g. $\alpha = 2.5$ and $\sigma = 0.25$ can be a good choice and leads to a water effect.

We noticed that applying piecewise affine distortion to the original frames was quite slow and we therefore ran most of our experiments without considering this technique. Since some of these transformations often generate new pixels (especially affine transformations and for example during a translation to the left where pixels are then generated on the right), we decided to fill these pixels. For this purpose, we decided to use different methods to fill them. As most of the successive frames of the same video are likely to be similar, we chose to perform data augmentation every s frames where s represents the root square of the mean number of frames per video in the considered class ($s = \sqrt{f}$). *Figure* 5 and 6 respectively show examples of some of the soft and hard augmentations we have produced.

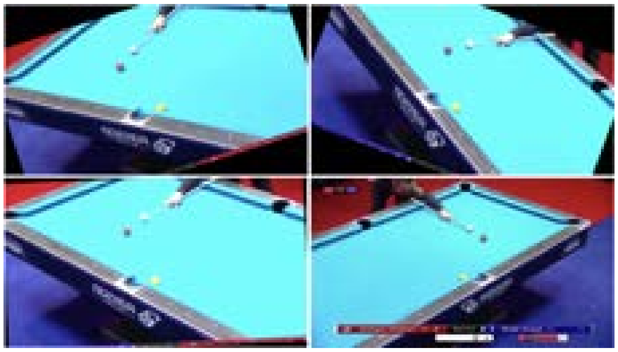


Figure 5: Data augmentation (smooth)

We thought about testing more advanced data augmentation methods but preferred to focus on the rest of our project due to time constraints. Among the methods we thought of are generative adversarial networks to generate new image samples and neural style transfer to combine the content of one image with the style of another.

C. Model Architecture

In order to quickly obtain decent models, we decided to first work on classical machine learning methods known for their efficiency on image recognition and particularly violence recognition tasks [**?**]. However, we did not spend much time on these models because we believe CNN-based neural architectures are much more suitable. We therefore decided to continue with these models. We describe below the models we tested.

1. XGBoost: We started with a basic learning model, namely the Extreme Gradient Boosting Classifier which is a boosting algorithm based on gradient boosted decision trees. Among its advantages, it applies better regularization techniques to reduce overfitting.

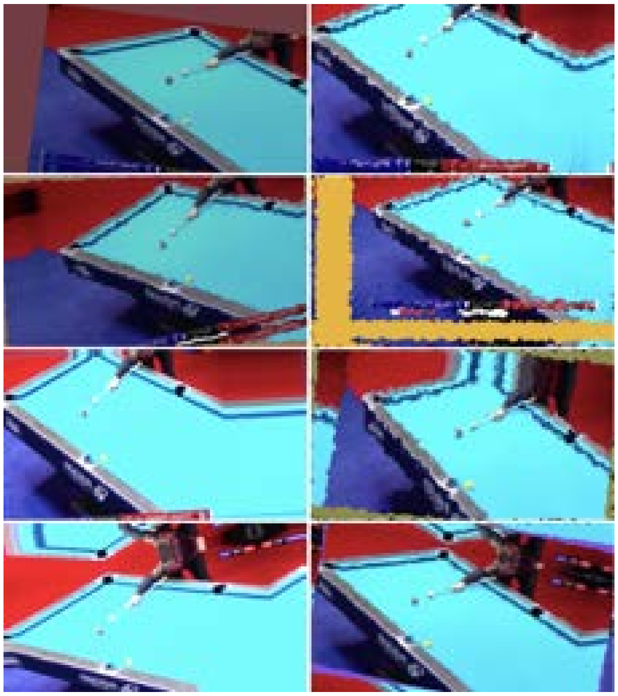


Figure 6: Data augmentation (hard)

2. Hybrid CNN-SVM: The proposed hybrid model combines the key properties of both classifiers. CNN works as an automatic feature extractor and SVM works as a binary classifier. The receptive field of CNN helps in automatically extracting the most distinguishable features from the videos frames. To convert the standard CNN into SVM we add a l2 norm kernel regularizer and pass a linear activation function in the final output layer. This way we create a hyperplane decision boundary between the two classes in order to separate them.

3. MobileNet: From the information we have gathered, MobileNet seems to have become a flexible solution for many moving object detection and image classification tasks. In particular, it seems to have already proven its performance in violence detection tasks, making it a strong candidate for our study. This is one of the first CNN architecture that is easily deployable on mobile applications. One of the main innovations is depthwise separable convolutions, which separates a classical convolution kernel into two kernels and thus reduces the number of operations required to perform the convolution. The second version of MobileNet introduces inverted residuals and linear bottlenecks to further improve performance.

4. Inception: Inception networks have shown excellent results in image classification tasks. They use a lot of tricks to push performance, both in terms of speed and accuracy. In particular, inception modules are used in CNNs to allow for more efficient computation and deeper networks through a dimensionality reduction with stacked 1×1 convolutions. The modules were designed to solve the problem of computational expense, as well as overfitting, among other issues. The solution is to take multiple kernel filter sizes within the CNN, and rather than stacking them sequentially, ordering them to operate on the same level.

IV. Experiments

A. Experimental setup

In order to obtain our dataset, we start by augmenting the data as described previously and match each frame with the nature of the video, i.e. either violent or safe.

We have chosen to use identical training parameters from one neural architecture to another in order to compare them fairly. In particular, we train our models over 100 epochs using binary cross-entropy loss with Adam optimizer [**?**], a learning rate of 0.0001 and mini-batches of 32 instances. We also decided to use additional callbacks to have more control over the training process. This includes stopping the learning process when the difference in validation loss drops below 0.005 for five successive epochs in order to prevent overfitting, adjusting the learning rates according to the current epoch as well as according to the improvement of the validation loss over five successive epochs, and finally saving the model after each epoch without overwriting the last best recorded model.

We believe that neural architectures could potentially lead to better results and, since we are limited in time, we decided to leave the fine-tuning of machine learning models aside.

B. Evaluation metrics

The augmented dataset we are working on is perfectly balanced, with about 50% of the training and test set frames coming from violent videos. It is important to note that because of this, we can rely on the accuracy of our different models, which is not always the case. Precision and recall are also important factors to consider. Depending on the objective, we may choose to place greater emphasis either on recall or precision.

If we assume, for example, that this project is carried out to detect harmful content on social medias, then the objective would be to limit the spread of violent content as much as possible. In this case, we would prefer a model that correctly classifies almost all potentially violent videos in order to filter them all out at the cost of some false positives, hence we would be more interested in high recall. At the same time, blacklisting videos that are considered violent when they are not can also greatly harm user experience, therefore the importance of precision should not be underestimated.

In this study, we do not know what the precise objective is and for this reason we will consider precision and recall equally. For this we will use their harmonic mean f1-score. Still, we have chosen to work with each important measure for the rest of the study in order to have a better view of our models (accuracy, precision, recall, f1-score as well as area under the ROC curve).

C. Model Training

Figures 7 and 8 illustrate the performances of the models along their training. These are the models that obtain the best performances as shown in *Table* 3. However, we observe that Inception achieves higher results and much faster, which makes it a particularly interesting candidate for violence detection. Moreover, we observe that some of theses models have not finished learning meaning that they should be able to achieve even better results over time.

D. Quantitative results

The best results were achieved by the XGBoostClassifier, MobileNet and Inception models. Their performance on the test set are shown in Appendix. More complete views of our results are provided at *Table* 2 and in 3 in Appendix.

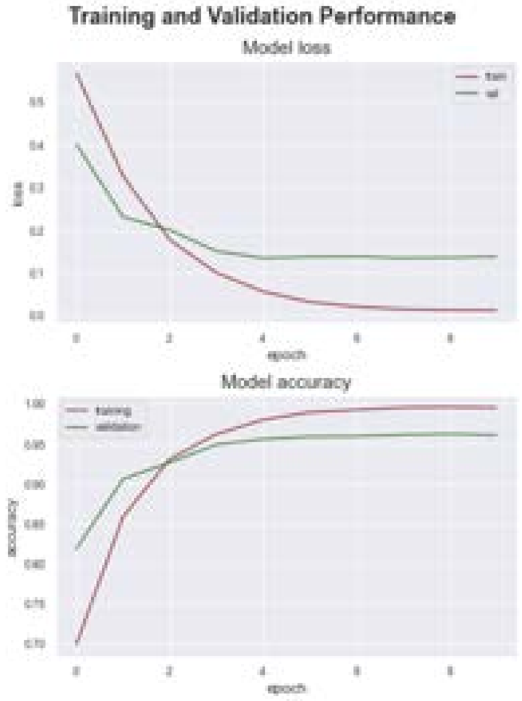


Figure 7: Training performance of Inception.

Method	**Acc**	**Prec**	**Rec**	**F1**	**AUC**
XGBoost	0.88	0.90	0.86	0.88	0.88
CNNSVM	0.74	0.77	0.70	0.73	0.74
Inception	**0.96**	**0.95**	**0.96**	**0.96**	**0.96**
MobileNet	0.92	0.92	0.93	0.92	0.92
MobileNet	0.84	0.84	0.84	0.84	0.84

Table 2: Comparing the proposed methods results on the RLVS [**?**] dataset.

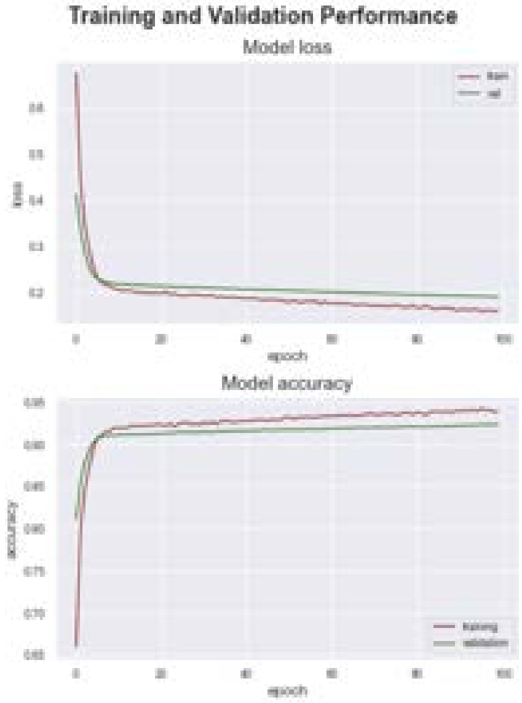


Figure 8: Training performance of MobileNet.

As a result of this project, we have obtained some good results using some standard machine learning techniques. In particular, the quality of predictions obtained by XGBoostClassifier is high enough to be satisfied. For example, we obtain a recall of 0.86 without even refining the model. Consequently, we believe that this model can be an excellent solution and hope to achieve better performance once refined. Using a hybrid CNN-SVM method unfortunately did not bring much and led to the worst results with a recall that does not exceed 0.70. As we would expect, methods based on pre-trained deep learning models like MobileNetV2 and InceptionV3 outperformed the more traditional approaches. For example, they were able to achieve a precision and recall of 0.95 and 0.96 respectively. However, while their performance is close, InceptionV3 achieved these results much faster than MobileNetV2.

E. Qualitative results

When we focus on the best performing model, we notice that in addition to achieving very good performances, it is able to classify some non-violent frames within a sequence labelled as violent. For example, on violent video V780 two boxers are violently fighting in the first frames, one of them is pushed aside by the referee who stops the fight in the subsequent frames. Our model manages to classify these first frames as violent until the referee separates the boxers, and the last frames are classified as safe except for one of them. In addition to this classification we also note the sharp decrease in the probability (from 0.99 to 0.26) that each frame is violent.

V. Conslusion

In this short project, we studied the detection of violent visual content in videos. For this we used the Real Life Violence Situations Dataset. We trained different models, including XBGBoost classifier and neural architectures, some of which were hybrid CNN-SVM models and others based on MobileNet and Inception. The latter proved to be the most efficient, especially and allowed us to achieve excellent results.

In this work we did not consider the audio data which should provide valuable additional information. Moreover we did not study the sequential aspect of the videos and did not add any recurrent part in our models. We also discussed the problems related to the dataset which contains only moderately violent videos, meaning that the models may not detect some very violent scenes such as wars.

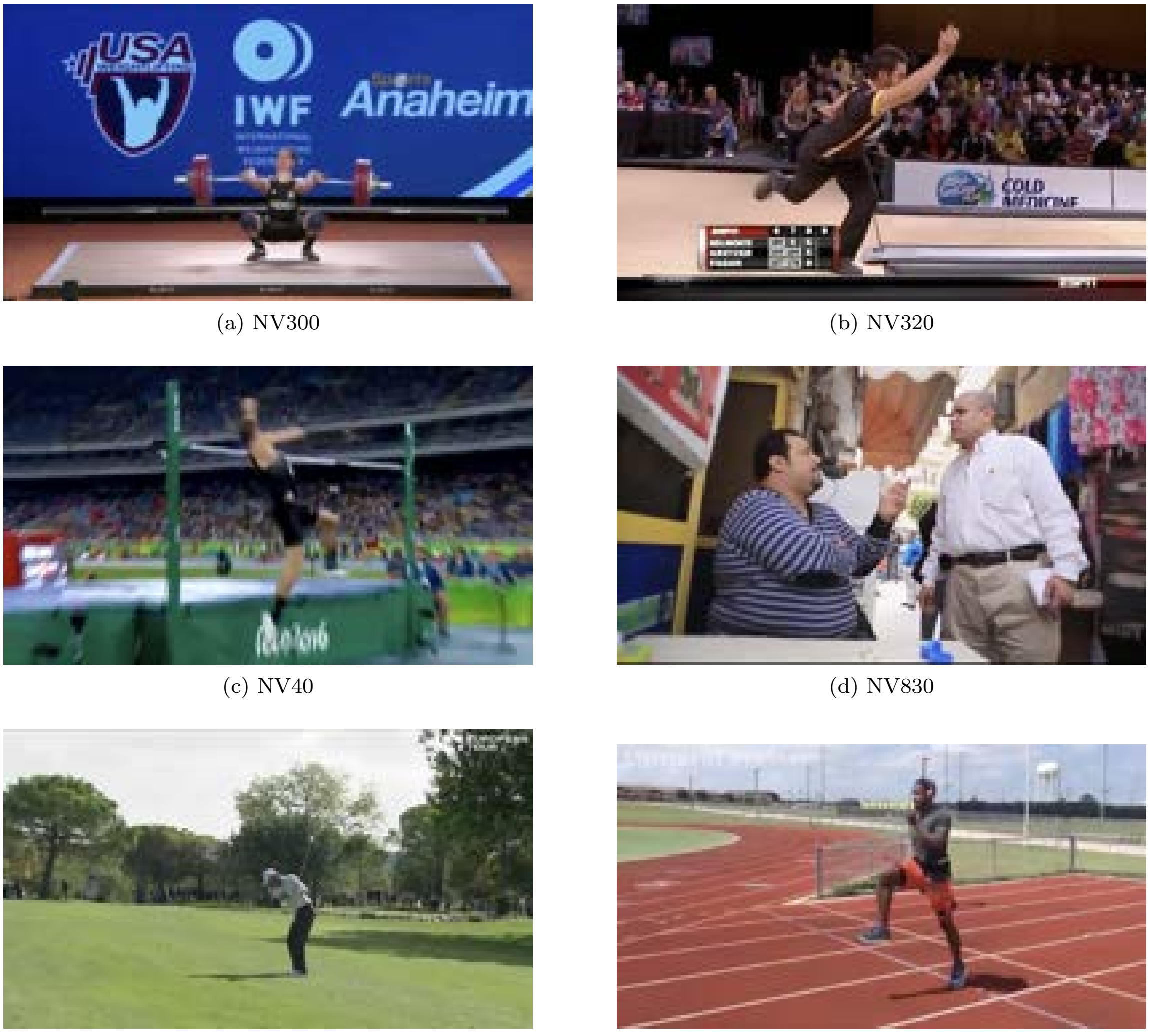
These subjects were not the focus of this competition, but we believe that further work on these matters could lead to reasonable solutions.

Appendix

Figure 9: Mosaic obtained from 200 violent videos.

Method	Accuracy	Precision	Recall	F1-Score	AUC	support
XGBoostClassifier	0.8836	0.8987	0.8638	0.8809	0.8836	(17808, 12288)
Hybrid CNN-SVM	0.7435	0.7666	0.6975	0.7304	0.7433	(17808, 64, 64, 3)
InceptionV3	**0.9555**	**0.9497**	**0.9617**	**0.9556**	**0.9555**	(17808, 128, 128, 3)
MobileNetV2	0.9234	0.9195	0.9274	0.9235	0.9234	(17808, 128, 128, 3)
MobileNetV2	0.8448	0.8431	0.8458	0.8445	0.8448	(17808, 64, 64, 3)

Table 3: Comparing the proposed methods results on the RLVS [**?**] dataset.



(a) NV300

(b) NV320

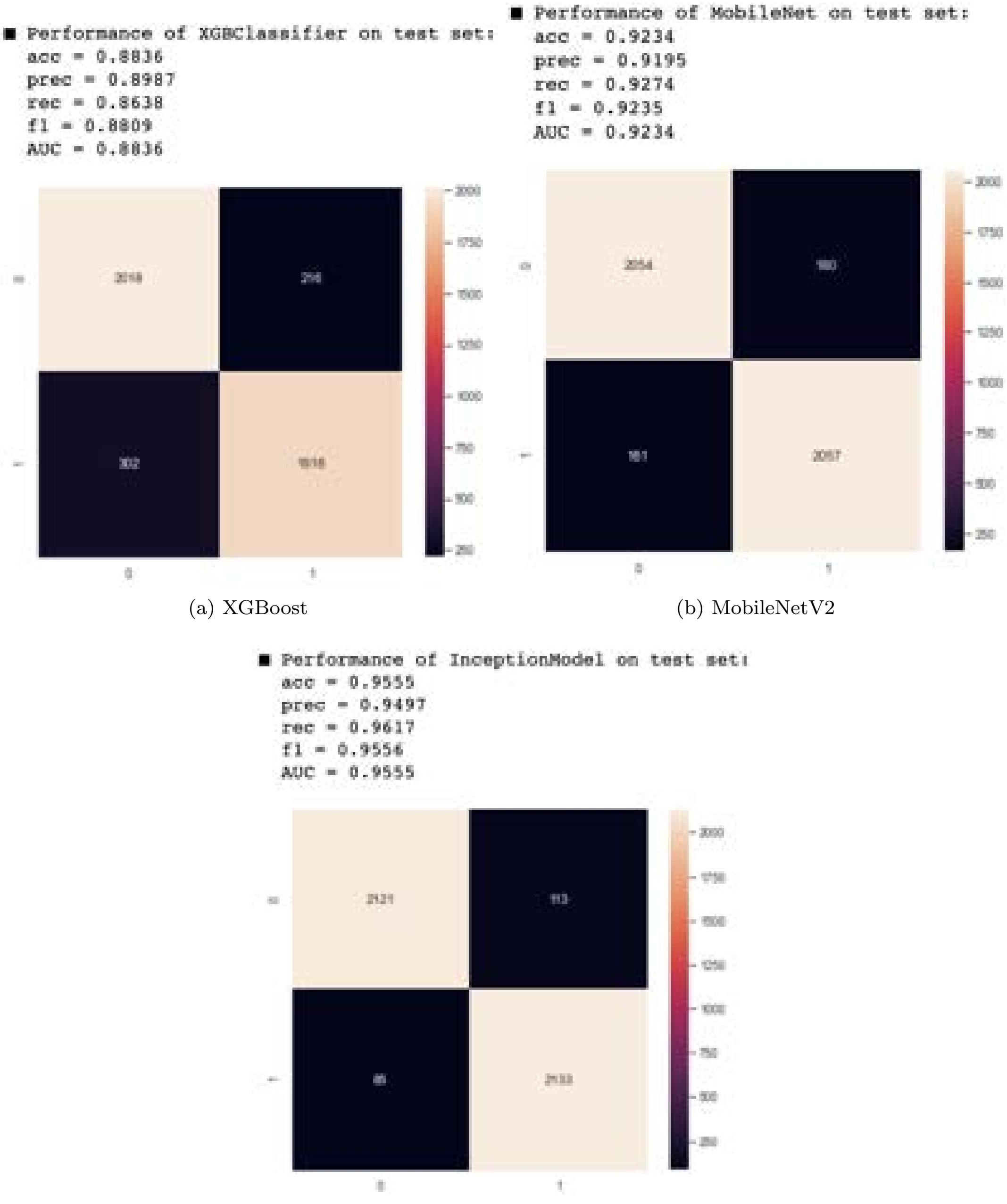
(c) NV40

(d) NV830

(e) NV851

(f) NV21

Figure 10: Example of non-violent videos.



```
■ Performance of XGBClassifier on test set:
  acc = 0.8836
  prec = 0.8987
  rec = 0.8638
  f1 = 0.8809
  AUC = 0.8836
```

(a) XGBoost

```
■ Performance of MobileNet on test set:
  acc = 0.9234
  prec = 0.9195
  rec = 0.9274
  f1 = 0.9235
  AUC = 0.9234
```

(b) MobileNetV2

```
■ Performance of InceptionModel on test set:
  acc = 0.9555
  prec = 0.9497
  rec = 0.9617
  f1 = 0.9556
  AUC = 0.9555
```

(c) InceptionV3

Figure 11: Performance of the best performing models on test set.

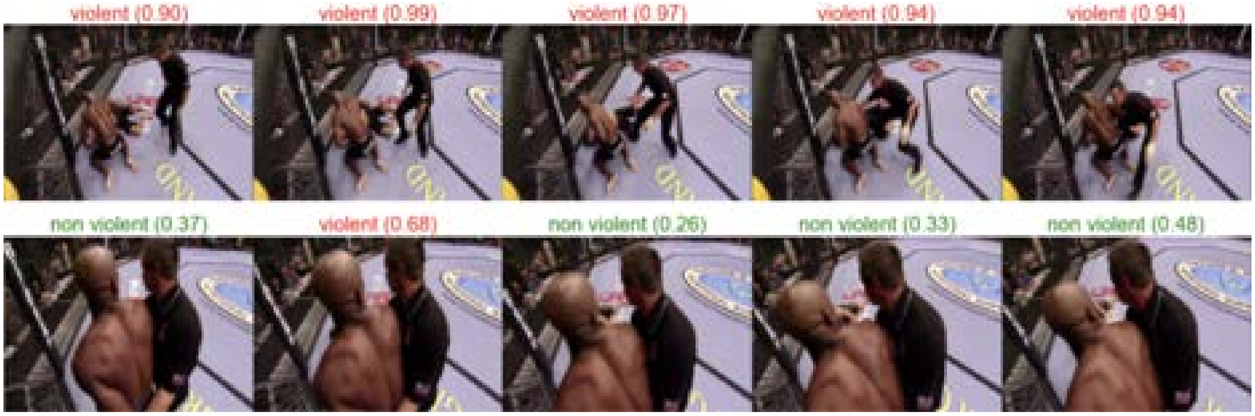


Figure 12: Frame by frame results of the best performing model (V760 video)