110K Sensitive Video Dataset (110k-SVD) Datasheet

I. Motivation For Dataset Creation

A. Why was the dataset created? (e.g., was there a specific task in mind? was there a specific gap that needed to be filled?)

We define sensitive video as any video that contains pornography or extremely violent scenes (that usually include but not limited to appearance of blood). At the time of creation we could not find any open access datasets that had this or similar definition and had more than 10.000 videos. This dataset was designed to be used specifically in the binary classification of entire videos, to determine whether a video contains sensitive content or not.

B. Has the dataset been used already? If so, where are the results so others can compare (e.g., links to published papers)?

No, the results of the baselines of this task are to be published.

C. What (other) tasks could the dataset be used for?

Tasks that include binary or multi label classification of the macro and micro classes of this dataset. Such as:

- Multi label classification (or tagging) of pornographic videos;
- Multi label classification of "Safe" (Videos that do not contain sensitive content);
- Binary classification of extremely violent videos (hereby refered as gore);
- Binary classification of pornography.

D. Who funded the creation dataset?

The creation of the 110K Sensitive Video Dataset database was supported by a joint challenge by Microsoft and Brazil's National Research Net (RNP) in 2019.

II. Dataset Composition

A. What are the instances?(that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

Each instance is a video (min 5 seconds, max 31 minutes).

TABLE I
General statistics of the two main classes of the dataset

	Sensitive	Safe
Video Count	67424	59651
Total Duration	6953:27:41	4852:53:31
Mean Duration	00:06:11	00:04:52
STD Duration	00:04:12	00:03:26
Max Duration	00:30:55	00:30:55
Min Duration	00:00:05	00:00:05
Total Size	1.2TiB	2.2TiB
Mean Size	19.3MiB	39.0MiB
STD Size	35.4MiB	42.3MiB
Features Size	519.4GiB	376.8GiB
Tag coverage	65392	51011
Tag coverage (%)	96,9862	85,5157

B. How many instances are there in total (of each type, if appropriate)?

As shown in Table II, it is divided into 59,651 safe videos and 67,424 videos with sensitive content. Those sensitive videos being 54,549 Pornographic Videos and 2,356 Gore Videos. Tag coverage refers to main tag annotation existence (videos also may have sub tags but no main tag).

C. What data does each instance consist of ? "Raw" data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution?

Each video will be distributed as extracted and processed visual and audio features. Each video file is associated with a id, a label, and a sequence size. There are people in the videos, but subpopulations are not identified.

D. Is there a label or target associated with each instance? If so, please provide a description.

Each video file is associated with a label (proper/improper) and id. Some examples of video data associated with the features: improper_29024487, proper_MqnZqzAxQTk, improper_gore122. There is also a main dataframe, this dataframe is indexed by video id and contains all the other gathered data, such as tags, subtags, file size, duration in seconds and title.

E. Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

For the main labels (sensitive/improper and safe/proper) there are no missing labels. For tags an sub tags, there is some missing information because either the website did not a tag system or the video had no tags on the website. The coverage of tags is shown in Table II.

F. Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

A individual might appear in multiple videos. This relationship was not collected and registered in the dataset. Other than this, there are no known relationships between instances.

G. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

This dataset is a sample, not necessarily random, of instances from a larger set. The larger set is all the videos from each crawled website. This dataset is representative of the sites we crawled because of the equally sampled variety of video types inside those sites and because of its large amount of instances.

However, the dataset is not a fully representative set of the entirety of videos on the internet, to be more representative of our definition of sensitive videos, the data should has to be collected from more video hosting sites. The sites used for gathering the videos were the most easily obtainable data at the time.

H. Are there recommended data splits or evaluation measures? (e.g.,training, development, testing; accuracy/AUC)

The instances come bundled in .npz files, each file represents a batch. We split our dataset into training/validation and testing batches. We publish features for all batches, but only publish labels for the train/validate batches. The user is free to split the training/validation as desired. For training evaluation, we recommend 20-fold cross validation to perform training and validation.

To report performance in the binary classification of sensitive videos, we recommend Precision (P), Recall (R) and, most importantly, the weighted F2-score. In this section we present a contextualized explanation these metrics.

In the context of sensitive content detection, *true positives* are videos predicted as sensitive and are in fact, sensitive. Likewise, *true negatives* are videos predicted as safe and are indeed safe. *False positives* are videos predicted as sensitive, but were safe, the same goes for *false negatives*, which are videos that were predicted as safe, but were predicted as sensitive.

Precision (Equation 1) measures how many videos predicted as sensitive (both true positives and false positives) are truly sensitive. The Recall (Equation 2) measures how many truly positive videos were correctly identified.

$$P = \frac{TP}{TP + FP} \quad (1) \qquad R = \frac{TP}{TP + FN} \quad (2)$$

Where TP, TN, FP, and FN denote the examples that are true positives, true negatives, false-positives, and false negatives, respectively.

$$F_\beta = (1 + \beta^2) \times \frac{P \times R}{(\beta^2 \times P) + R} \quad (3)$$

The F_β-score, defined in Equation 3, evaluates the classifier by the harmonic mean between Precision and Recall. To account for label imbalance, after calculating the F2-score metrics for each label, we find their average weighted by support (the number of true instances for each label).

While the F1-score represents an balanced performance metric, the F2-score gives twice more weight to the recall than to precision, which means that the metric is more focused on the recall of a solution.

We chose the weighted F2-score as our main evaluation metric because when detecting sensitive content it is more important to predict a truly sensitive video than to predict a safe video as sensitive.

I. Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

There might be annotation divergences in the tags and sub tags of the pornography videos, since the videos were tagged by users and not by a centralized annotation group. We can not guarantee that frames and/or audio clips do not appear in other videos since there was no direct contact with the videos during dataset creation. There was however a duplicate removal step in the creation of the dataset, detailed later in this document.

J. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

Everything needed to perform the proposed tasks is included.

Any other comments?

III. Collection Process

A. What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

There was no direct human curation, the videos were automatically collected based on their titles and tags. We created crawlers to automatically collect the videos.

B. How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

If available, the title, tags, and sub tags of the video were collected and stored by the crawler. The video feature extraction process was already validated for multi label video classification [1]. To validate the feature extraction process for the task we propose we also trained and tested baseline models and archived an F2-score of 99% in our test sub set and 88.83% in a popular pornography dataset [2].

C. If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The sampling strategy was to collect at least the amount of the less numerous tag for each tag, then, to complete the amount of collected videos the sampling probabilities were proportional to the database distribution of each tag.

D. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Two graduate students.

E. Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The dataset was crawled from january of 2019 to december of 2019, this timeframe does not match the creation of the data associated with the instances.

IV. Data Preprocessing

A. Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Once the raw data was crawled, we performed feature extraction in all videos successfully collected. To generate image frame features and audio features we decode each video at approximately 1 frame-per-second and feed an InceptionV3 network [3] pre-trained on the ImageNet[1] dataset.

We also make use of a AudioVGG [4] network with pre-trained weights in the Audioset[2] dataset to extract the audio embeddings. Each of these CNNs were used as published by their authors; the only modification was the removal of classification layers in both CNNs to obtain their respective embeddings.

Next, we apply Principal Component Analysis (PCA) [5] to each of the outputs to reduce the dimensions of both embeddings and to generate feature vectors of size 1024 and 128 for frame and audio embeddings respectively.

We concatenate both image and audio embeddings extracted in the current frame and audio window in order to compose the final embeddings as a sequence of the same size of the number of seconds of the video. After this concatenation, each time-step has 1,152 features: 128 audio features and 1024 frame features.

Notice that with this approach, the video is transformed into a time series, and to use it in non-sequential models (*e.g.* SVM, KNN, and MLP) we need to turn this sequence into a single feature vector that represents the whole video. In our setting, we did that by taking the average, median, standard deviation, min, and max values for each feature to represent the entire video. In summary, we turn the sequence of features with size n and shape n by 1,152 into a single feature with shape 1 by 5,760.

We also filtered out short and long videos. For the short videos, we defined that the minimun lenght of a video was 5 seconds based on [2], which were 0.09% of the dataset. To define the maximum lenght of a video in the dataset, first we removed all videos with less than 5 seconds, then we calculated the mean and standard deviation of each video's duration. The maximum lenght of a video in the dataset is $mean + 2 * std$, which resulted in approximately 31 minutes and covered 98,94% of the videos.

Not all video's features were successfully extracted for multiple reasons, such as, corrupt data, unkown format, and missing audio. For those videos with missing audio or image, the features were still generated, but their respective modal feature were zeros. Those videos which do not had any features successfully extracted were removed from the dataset.

We also removed any duplicated videos that were detected, for duplicate video detection we used, we matched either id,

[1] http://www.image-net.org/
[2] https://research.google.com/audioset/

title or checksum.

We recommend equally balancing both main labels (sensitive/improper and safe/proper), so that both main classes have the same number of instances. One could also choose not to balance both classes equally, since our main metric already take label imbalance into account. Additionally, when removing excess sensitive content (while balancing), we recommend removing only pornography videos in order to not lower the amount of gore videos.

B. Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

We can not provide a link for the raw video data, but we are open to include other feature extraction methods. If there are any suggestions for better or newer feature extraction methods, please, get in contact with us.

C. Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the feature extraction method and our preprocessing is available in the github repository: `https://github.com/TeleMidia/Sensitive-Video-Dataset`.

D. Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? If not, what are the limitations?

Although our baseline tests perform well on the gore detection task, there is still a relative small amount of gore videos in our dataset. Furthermore, the is no manually curated dataset comparative to the gore videos. Mainly because of the difference in our definition of violence, which is just highly violent scenes such as death, mutilation and torture.

E. Any other comments

V. Dataset Distribution

A. How will the dataset be distributed? (e.g., tarball on website, API, GitHub; does the data have a DOI and is it archived redundantly?)

The dataset scripts, updates, papers, and addictional information will be hosted on github: `https://github.com/TeleMidia/Sensitive-Video-Dataset`. The dataset itself will be hosted by the IEEE Dataport:

- DOI: 10.21227/sx01-1p81
- URL:`https://ieee-dataport.org/documents/sensitive-video-dataset`

The dataset will be distributed in multiple .npz files, organized in multiple directories:

- *train_val_batches*
- *test_subset_batches*
- *non_sequential_train_val_batches*
- *non_sequential_test_subset_batches*

There is also a main dataframe, this dataframe is indexed by video id and contains all the other gathered data, such as tags, subtags, file size, duration in seconds and title.

Each npz file represents a batch of variable size, but all split to have at max 4 Gbs when loaded to memory. Each npz file has keys and values, the keys are string in the format ¡label¿_¡video id¿. Some examples of keys in the npz file: "improper_29024487", "proper_MqnZqzAxQTk", "improper_gore122".

The values are the videos features stored in numpy arrays, of varying shape, depending on the dataset variation (sequential or non-sequential).

The dataset has two variations:

- Sequential: Each sample remains as it was extracted, a single video generates a sequence of N samples. In this variation, inside each npz file each instance is represented by a N by 1152 numpy array.
- Non-Sequential: All samples of a video are aggregated into a single sample, resulting in each instance having a shape of 1 by 5760, this single sample summarizes the entire video.

The data is archived redundantly.

B. When will the dataset be released/first distributed? What license (if any) is it distributed under?

It is available on IEEE Dataport (`https://ieee-dataport.org/documents/sensitive-video-dataset`) under Ceative Commons Attribution 4.0 International (CC BY 4.0).

C. Are there any copyrights on the data?

No.

D. Are there any fees or access/export restrictions?

There are no fees or restrictions.

E. Any other comments?

VI. Dataset Maintenance

A. Who is supporting/hosting/maintaining the dataset?

The dataset is hosted by IEEE Dataport and maintained by the authors.

B. Will the dataset be updated? If so, how often and by whom?

There are no expected updates on this dataset.

C. How will updates be communicated? (e.g., mailing list, GitHub)

If any, updates will be comunicated via the dataset's GitHub page/repository.

D. If the dataset becomes obsolete how will this be communicated?

Through the dataset's GitHub page/repository.

E. Is there a repository to link to any/all papers/systems that use this dataset?

The links about papers and works using our dataset will be held on the dataset's github repository:`https://github.com/TeleMidia/Sensitive-Video-Dataset`.

F. If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?

Others are free to use and modify our datasets. Contributions can be discussed via email (pedropva@telemidia.puc-rio.br).

VII. Legal and Ethical Considerations

A. Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No. The data was crawled from public web sites. The reproducible videos were not assessed by anyone and will not be distributed, only the features will be distributed. Those features can not be reverted or recreated into reproducible videos.

B. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

No, all videos were crawled from public sources, furthermore, only video the features will be distributed. Those features can not be reverted or recreated into reproducible videos.

C. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why

No, only video the features will be distributed. Those features can not be reverted or recreated into reproducible videos.

D. Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No, only video the features will be distributed. Those features can not be reverted or recreated into reproducible videos.

E. Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

Not applicable.

F. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

Not applicable.

G. Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

Not applicable.

H. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

Not applicable.

I. Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Not applicable.

J. Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

Not applicable.

K. If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

Not applicable.

L. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis)been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

Not applicable.

M. Any other comments?

REFERENCES

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

[2] Daniel Moreira, Sandra Avila, Mauricio Perez, Daniel Moraes, Vanessa Testoni, Eduardo Valle, Siome Goldenstein, and Anderson Rocha. Pornography classification: The hidden clues in video space–time. *Forensic science international*, 268:46–61, 2016.

[3] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[4] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.

[5] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.