Sensitive-Video Analysis

Daniel Moreira[1], **Siome Goldenstein**[1], **Anderson Rocha**[1]

[1]Instituto de Computação – Universidade Estadual de Campinas (Unicamp)
Campinas – SP – Brazil

{daniel.moreira, siome, anderson.rocha}@ic.unicamp.br

Abstract. *Sensitive videos that may be inadequate to some audiences (e.g., pornography and violence, towards underages) are constantly being shared over the Internet. Employing humans for filtering them is daunting. The huge amount of data and the tediousness of the task ask for computer-aided sensitive video-analysis, which we tackle in two ways. In the first one (sensitive-video classification), we explore efficient methods to decide whether or not a video contains sensitive material. In the second one (sensitive-content localization), we explore manners to find the moments a video starts and ceases to display sensitive content. Hypotheses are stated and validated, leading to contributions (papers, dataset, and patents) in the fields of Digital Forensics and Computer Vision.*

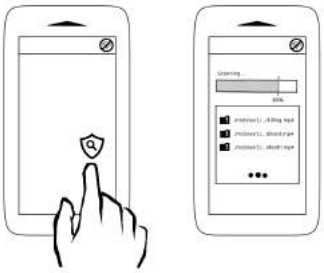
1. Introduction

We can define sensitive video as every motion picture whose content may pose threats to inappropriate audiences (e.g., children or unwary spectators). Typical representatives include scenes depicting pornography, violence, animal cruelty, child abuse, etc.

Due to the present easiness and multitude of ways to produce, share, and send video streams over the Internet, the diversity of content is untold. Within such diversity, it is not hard to imagine that some streams may be sensitive, thus demanding somebody to filter them. However, the employment of humans to frequently analyze large troves of sensitive data often leads to stress and trauma, justifying the research for computer-aided analysis, for alleviating the job of moderators.

Notwithstanding, automated sensitive-video analysis is a challenging problem. Besides the increasingly large amount of data to be analyzed, nowadays, video playing and online live streaming happen mostly on mobile devices, practically anywhere, in myriad of potentially-inappropriate situations (e.g., at work, schools, etc.). How to design ubiquitous and efficient solutions, which can operate on the consumer side, even on devices with limited hardware is the question of the day. Moreover, there are situations in which the detection of sensitive content is urgent. That happens, for example, in Forensic scenarios, in which the fast identification of child pornography, among millions of files, may allow catching red-handed criminals. In addition to such challenges, the literature of sensitive-video analysis has consistently reported that the incorporation of video motion information improves the final system effectiveness, at the expense of demanding high computational power. As a consequence, the performance is usually impaired, specially in terms of memory footprint and runtime.

Observations and questions such as the aforementioned ones have driven us toward the statement of the following research hypotheses, which guided the development of the Ph.D. thesis at hand:



(a) (b)

Figure 1. Application example of sensitive-video classification in a smartphone. In (a), the user activates a scanning app for sensitive content, and in (b), the app enlists the sensitive (e.g., violent) videos, with a progress bar depicting the scanning progress.

H1 It is possible to efficiently use video temporal information for effective sensitive-content classification, regarding low-memory footprint[1] and small processing time[2], by combining simplified space-temporal video interest-point detection and description, with entire-footage representation through a single feature vector.

H2 It is possible to localize sensitive content within the video timeline by means of the classification and fusion of time-overlapping video snippets[3].

As one might observe, for the sake of research scope definition, we tackle the problem of sensitive-video analysis as either (i) a problem of classifying sensitive video content, or (ii) a problem of localizing sensitive content within the video timeline. As a direct result of that, the present work is divided into two parts. In Section 2 we explain the *Part I* of the research, which is respective to *Sensitive-Video Classification*, while in Section 3, we focus on the *Part II*, which is related to *Sensitive-Content Localization*. Regardless of such division, in Section 4, we summarize the accomplishments of the research, while conclusions and future work are discussed in Section 5.

2. Sensitive-Video Classification

Sensitive-video classification is the decision problem of determining whether or not a given video stream has *any* occurrence of a particular sensitive content. Figure 1 depicts a possible application of a sensitive-video classifier.

Focusing on the task at hand, and keeping hypothesis *H1* in mind, we introduce an end-to-end pipeline for motion-aware sensitive-video classification, which is designed to be efficient (i.e., to be fast and to present low-memory footprint). Such pipeline consists of a three-level Bag-of-Visual-Words (BoVW) -inspired solution, which efficiently employs temporal information as an effective discriminative clue for video classification. It incorporates temporal information in the low and mid levels, by means of a novel space-temporal interest point detector and video descriptor — namely Temporal Robust Features (TRoF) — and entire-footage mid-level feature pooling, respectively. It relies on Gaussian-Mixture-Models (GMM)-based codebooks, Fisher Vectors, and a linear Support Vector Machine (SVM), one of the most effective combinations that were ever reported in the BoVW-related literature. It is of general purpose, in the sense that it can be used — without step modifications — for the detection of diverse sensitive content types (e.g., gore scenes, child abuse, cruelty to animals, etc.) thus broadening the scope and applications of the present research.

[1]We consider that a solution has low-memory footprint, if it is at least amenable to deployment on contemporary (as of 2015 and 2016) mobile devices, such as smartphones and tablets.

[2]Preferably close to real time, i.e., 24-30 frames per second.

[3]A snippet is any video excerpt, with arbitrary length.

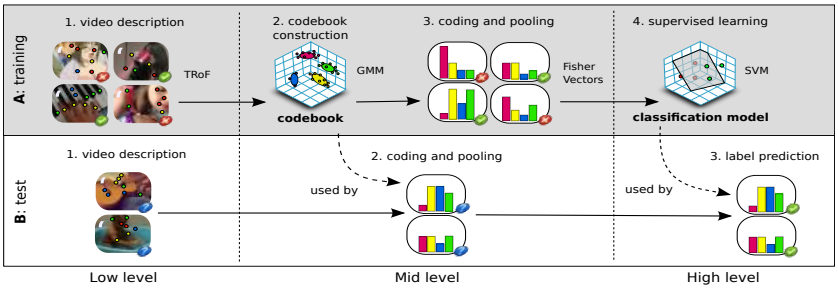


Figure 2. The three-level proposed pipeline for sensitive-video classification. On the top, the darker box depicts the training phase, in which the system is fed with labeled videos. On the bottom, the lighter box depicts the test phase, in which the system predicts the label of arbitrary videos.

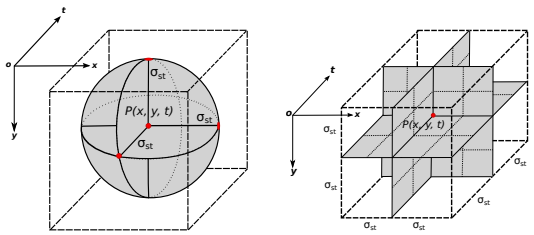


Figure 3. TRoF voxel samplings. The gray areas refer to the described voxels, which are arranged around position $P\ x, y, t$, obeying a spatio-temporal scale of $2\sigma_{st}$. P and σ_{st} come from a formerly detected interest point, and refer to its position and interest scale. (a) SURF-based sampling. (b) HOG-based sampling.

Figure 2 depicts the proposed pipeline, with the inherent three levels. For a more detailed representation, please refer to Figure 3.1 in the original text. Within such pipeline, efficiency mainly relies upon the use of TRoF: an interest point detector and video descriptor that accounts for a lightweight space-temporal alternative, when compared to the more computationally intensive space-temporal solutions from the literature. It is fast and presents low-memory footprint, what makes it amenable to run on limited hardware, such as mobile devices. To reach such efficiency, TRoF relies on a sparse strategy, which detects an optimized amount of space-temporal interest points within the video timeline. The detection process is Hessian-based, and is underpinned by the concepts of integral videos and box filters for fast computation. The description process, in turn, is optimized by selecting only a small amount of video voxels around the previously detected interest points.

Figure 3 depicts the two explored voxel-sampling strategies within the TRoF video description process. For a complete explanation, please refer to Section 3.2 in the original text. TRoF detection capabilities can be watched through illustrative animated videos (please refer to videos [Moreira 2016, Moreira 2017] online).

We validate the TRoF-based pipeline for both pornographic and violent content classification. In summary, TRoF constitutes an interesting alternative for dealing with the effectiveness vs. efficiency tradeoff. By means of the performed experiments, we verify that, in spite of not being statistically different to state-of-the-art space-temporal video descriptors, the proposed solution can reach real-time performance, and presents the lowest memory footprint. In addition, with a classification accuracy of 93%, it remarkably outperforms four third-party off-the-shelf pornography detectors, including two commercial products, by providing an error reduction of over 45%. Results are published in detail through a journal [Moreira et al. 2016] (regarding pornography classification) and a conference paper [Moreira et al. 2017b] (regarding violence classification). According to such results, we found strong evidence that hypothesis *H1* is valid.

(a) (b) (c)

Figure 4. Application example of sensitive-content localization. In (a), the user starts to play a chosen video, within a tablet, through a safe video player. In (b), the video that is being played is about to show sensitive content (pornographic). In (c), the pornographic scenes are properly censored.

3. Sensitive-Content Localization

Sensitive-content localization is the search problem of finding sensitive scenes within a video timeline. A sensitive-content locator must return the time intervals a video stream starts and ceases to display sensitive content. Figure 4 depicts a possible application.

Focusing on the task at hand, and with hypothesis *H2* in mind, we design a novel high-level multimodal fusion pipeline for sensitive-content localization, which is based on the combination of different and independent sensitive-snippet classifiers. In summary, we recommend classifying the content of distinct time-overlapping snippets, in order to provide a dense sampling and a dense classification of the video timeline. For integrating the snippet classifiers, we introduce a new late fusion technique, which seamlessly combines the sensitiveness scores that are returned by each classifier. Scores that refer to the same video instant of interest are used to generate a single time-localized fusion feature vector. From the many generated time-localized fusion feature vectors, we learn which ones better represent sensitive and non-sensitive video moments, by employing machine-learning techniques (such as Naïve Bayes Classification and SVM). Given that such fusion vectors are composed of prior sensitiveness-classification scores, we essentially propose a meta-learning solution. Besides that, since each snippet classifier can freely rely on a particular data modality (e.g., still video frames, audio stream, video space-time, etc.), the proposed solution has an important multimodal capability.

Figure 5 depicts the training steps of the proposed pipeline for a particular toy case, with two snippet classifiers, and snippet alignments of three training movies. For the sake of illustration, the alignment of snippets is represented in activity 2 (*Snippet Alignment*), while the resulting fusion vectors are depicted in activity 3 (*Fusion Vector Extraction*). For a more detailed description, please refer to Chapter 6 in the original text.

We validate the solution for both pornographic and violent content localization, combining visual and auditory features. In the particular case of pornography, we fail at identifying around only five minutes in every hour of pornographic content, indicating a reasonable content filter. In the particular case of violence, the present pipeline has led us to reach second place in an international competition of violent scenes localization [Avila et al. 2014].

Figure 6 shows an impressive qualitative result of pornographic content localization over a 1.5-minute long video sample. Mislocalization is represented by black dots, which are only a few, around the instants of transition from non-pornographic (in white) to pornographic (in red) content, and vice-versa. Please refer to the original text for more easy and difficult examples.

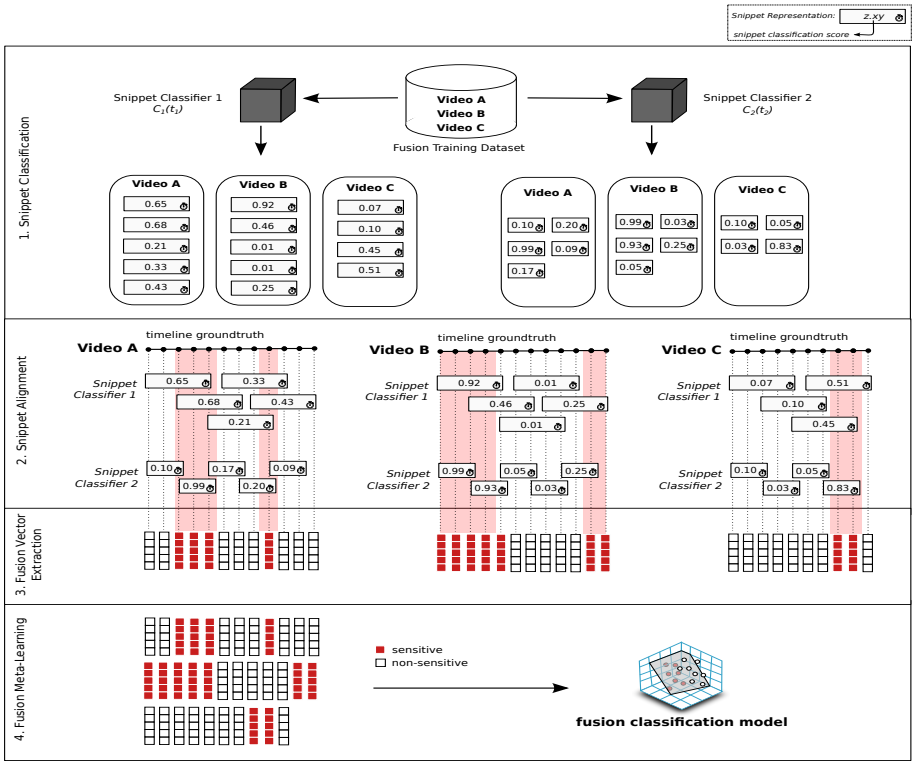


Figure 5. Toy-case example of the proposed fusion training pipeline. The method starts with the submission of *Videos A, B, and C* to two different snippet classifiers that need to be fused (classifiers C_1 t_1 and C_2 t_2). The method ends with a meta-learned fusion classification model, which must be stored for further use, during the test system operation.

Figure 6. Localization quality over a 1.5-minute long pornographic video sample. Red and white areas depict the localization groundtruth: red for positive, and white for negative. Black dots represent the mislocalization: the lesser their quantity, the better the result.

As an important additional contribution, we also introduce the *Pornography-2k* dataset, a large frame-level-annotated pornographic video dataset that is useful for pornographic content localization, since we provide frame-level annotation for its 140 hours of video footage. To the best of our knowledge, Pornography-2k is the first freely-available pornographic dataset in the literature that provides binary annotation (i.e., pornographic vs. non-pornographic) for every one of its frames.

Regardless of the target sensitive concept, the present pipeline was subject to the deposit of two patents, in partnership with Samsung Electronics, one in the Brazilian National Institute of Industrial Property (INPI) [Avila et al. 2016a], and the other in the United States Patent and Trademark Office (USPTO) [Avila et al. 2016b]. Finally, it was submitted to the scientific community's appreciation by means of a journal paper [Moreira et al. 2017a], yet under revision.

According to the results that are discussed in Chapters 7 and 8 of the original text, we found strong evidence that hypothesis *H2* is valid.

4. Research Accomplishments

In summary, the main accomplishments of this research are:

- Two patent filings, one deposited in INPI [Avila et al. 2016a], and the other deposited in USPTO [Avila et al. 2016b].
- Patent licensings to Samsung Electronics and subsequent technology transfer.
- Two journal publications, one already published [Moreira et al. 2016], and the other under review [Moreira et al. 2017a].
- Three conference papers [Avila et al. 2014, Moreira et al. 2015, Moreira et al. 2017b].
- Second-place award in an international competition of video violence localization [Avila et al. 2014].

5. Conclusions and Future Work

By verifying the stated hypotheses, the Ph.D. thesis at hand contributes to the fields of Digital Forensics and Computer Vision with the following novelties:

- End-to-end pipeline for efficient motion-aware sensitive-video classification.
- Space-temporal video interest point detector and video content descriptor (TRoF).
- High-level multimodal fusion pipeline for sensitive-content localization.
- Large frame-level-annotated pornographic video dataset.

The proposed solutions are generic enough to be straightforwardly applied for different types of sensitive content. For validation, experiments are conducted for the *classification* of video *pornography* and of video *violence*, as well as for the *localization* of *pornographic* content, and of *violent* content, leading to interesting results. As future work, other types of sensitive content can be validated (e.g., child pornography), as well as the fusion with the now popular deep learning strategies can be investigated.

References

Avila, S., Moreira, D., Perez, M., Moraes, D., Cota, I., Testoni, V., Valle, E., Goldenstein, S., and Rocha, A. (2014). RECOD at MediaEval 2014: Violent Scenes Detection Task. In *MediaEval*, pages 1–2.

Avila, S., Moreira, D., Perez, M., Moraes, D., Testoni, V., Valle, E., Goldenstein, S., and Rocha, A. (2016a). Método Multimodal e em Tempo Real para Filtragem de Conteúdo Sensível. Patent BR 10 2016 007265 4.

Avila, S., Moreira, D., Perez, M., Moraes, D., Testoni, V., Valle, E., Goldenstein, S., and Rocha, A. (2016b). Multimodal and Real-Time Method for Filtering Sensitive Media. Patent US 15/198,626.

Moreira, D. (2016). `https://www.youtube.com/watch?v=ZfaW5kvXjMo`.

Moreira, D. (2017). `https://www.youtube.com/watch?v=yoV4b1CQ1aY`.

Moreira, D., Avila, S., Perez, M., Moraes, D., Testoni, V., Valle, E., Goldenstein, S., and Rocha, A. (2015). RECOD at MediaEval 2015: Affective Impact of Movies Task. In *MediaEval*, pages 1–2.

Moreira, D., Avila, S., Perez, M., Moraes, D., Testoni, V., Valle, E., Goldenstein, S., and Rocha, A. (2016). Pornography Classification: The Hidden Clues in Video Space-Time. *Elsevier Forensic Science International*, 268:46–61.

Moreira, D., Avila, S., Perez, M., Moraes, D., Testoni, V., Valle, E., Goldenstein, S., and Rocha, A. (2017a). Multimodal Data Fusion for Sensitive Scene Localization. *Elsevier Information Fusion (under review)*.

Moreira, D., Avila, S., Perez, M., Moraes, D., Testoni, V., Valle, E., Goldenstein, S., and Rocha, A. (2017b). Temporal Robust Features for Violence Detection. In *IEEE WACV*, pages 392–399.