IJoFCS (2012) 2, 21-32

e-forensic PRESS

The International Journal of
FORENSIC COMPUTER SCIENCE

www.IJoFCS.org

DOI: 10.5769/J201202002 or http://dx.doi.org/10.5769/J201202002

Quick Identification of Child Pornography in Digital Videos

Mateus de Castro Polastro and Pedro Monteiro da Silva Eleuterio
Brazilian Federal Police
Campo Grande/MS
E-mails: polastro.mcp@dpf.gov.br, pedro.pmse@dpf.gov.br

Abstract - Computer forensics has the main objective to find digital evidences of crimes. One of the most researched digital crimes is the sexual abuse of children, including the production, sharing and possession of child pornographic files. Aiming to quickly detect files of child pornography at crime scenes, the NuDetective Forensic Tool was previously developed and it uses techniques like nudity detection in images and videos, among others. To automatically detect child pornography videos, a prior approach was developed, based on extraction and sampling a fixed number of frames of each video files, independently of the video duration. This work proposes a new adaptive sampling approach, considering the video duration, with the objective to increase the detection rate and/or reduce the runtime. Several experiments were performed and the results proved that this new approach is more appropriate to be used in the automatic detection of child pornographic videos at crime scenes, with detection rates around 87% and a reduction about 45% of the runtime over previous experiments. An experiment with a real forensic case was also performed and proved that the new approach can be used to quickly identify such illegal files at crime scenes.

Keywords - child pornography, nudity detection, video frame extraction, computer forensics, crime scene analysis.

I. Introduction

Nowadays, it is very easy to produce multimedia files such as images and videos in high resolution. Cameras and camcorders are typical equipments that most people have access and can use. However, some people use such equipments to produce illegal content, such as those related to child pornography, recording images and videos of child sexual abuse. With the ease of the Internet and the bandwidth increase [1], people around the world can share and have access to these illegal files, mainly through peer-to-peer networks [2], World Wide Web (WWW), encrypted private networks, such as GigaTribe[1], and attachments of electronic and instant messages.

1 GigaTribe is a confidential file sharing software - available at http://www.gigatribe.com

Paper submitted on November 8th, 2012.

The laws of many countries regulate that possession of child pornographic files is a heinous crime [3]. Countries such as Australia, Canada, United States, United Kingdom, and Brazil [4], have legislation to criminalize people who save these files on any type of digital storage device. Consequently, forensic analysis is being increasingly required to prove the existence of this type of content, including at crime scenes. Real forensic studies [5] and the experience of the authors show that the main difficulty in this type of analysis is to identify the files containing child and teen pornography among often millions of files that can be stored in these digital devices.

Moreover, the identification of these file types at crime scenes can bring great benefits such as immediate arrest of the pedophile and the correct selection of the material to be seized. To accomplish this identification at crime scenes, the forensic examiners have several challenges, for example, they do not have enough time to individually analyze millions of files and also do not have a robust computational infrastructure, since all the equipment used must be transported to the crime scene. To help the forensic examiners to quickly identify child pornographic files at crime scenes, the NuDetective Forensic Tool [6] was developed. NuDetective can quickly find the suspicious files that must be analyzed by the forensic examiners. The Tool performs pixel and shape analysis to automatic detect human nudity images (*Image Analysis*), including high resolution images [7], searches for file names that may contain typical expressions related to child and teen pornography (*FileName Analysis*), and also calculates the hash value of suspect files to compare to previously known hash list (*Hash Analysis*). Furthermore, in a previous work [8], the authors developed a strategy to also identify videos of child pornography (*Video Analysis*). The strategy was based on the extraction of video frames, using the nudity detection algorithms already implemented by *Image Analysis* feature. Through several experiments, the authors defined a Cutoff Point, which indicates whether or not the video will be classified as suspicious by NuDetective. The Tool detection rates are around 95% in images and 85% in videos, and many law enforcement agencies have been using the Tool at crime scenes.

Video files have important features that make quick automatic detection a hard task. Files are usually large and have different formats and encodings. Besides that, child pornographic videos are restricted material to be handled and may have low resolution. In the literature, several other strategies have been found to detect pornography in videos, such as keyframe analysis, motion detection, sliding window periodicity, the bag-of-visual-words, the bag-of-visual-features, periodicity detection, and motion histograms [9, 10]. However, none of them were specifically developed to detect child pornography, which have different characteristics compared to adult pornography. In general, these strategies are complex, perform many processing, and require a lot of time to be performed.

The video detection strategy implemented in NuDetective samples a fixed number of video frames, regardless of the video duration. Despite the great results achieved with that approach [8], the authors felt that using a fixed number of frames for all the videos may not be the best approach and may delay the detection process, especially in shorter videos.

In this scenario, the authors have studied ways to improve the *Video Analysis* feature provided by NuDetective, creating a new adaptive approach to automatically detect videos of child pornography. The new approach needs to maintain the previous requirements, as be quick enough to be performed at crime scenes together with good detection rates. However, it must use different frame sample sizes for different video durations, trying to attain better detection rates and/or runtimes. To achieve the objective, the authors performed several experiments with different sampling strategies, allowing the definition of a new formula that defines how the NuDetective will sample frames of child pornographic videos.

This paper is organized as follows: Section 2 presents some concepts of video digital files and the state-of-art of video pornography detection, a brief description of the main features of NuDetective Forensic Tool, and, also, a basic review of mathematical functions. Section 3 presents the proposed approach for sampling video frames. Section 4 shows the evaluation of the new approach, including the performed experiments and results analysis. Sections 5 and 6 present the conclusions and future work, respectively.

II. Background

In this section, key concepts related to this work are reviewed, as some concepts of a video file, the state-of-art of nudity and pornography detection in videos, and the main features of the NuDetective Forensic Tool, including its detection algorithms. This section also reviews the mathematical functions, important for understanding the choice of the new adaptive sampling approach.

A. Basic Video Concepts

Unlike digital still image, which is timeless and static, video is dynamic and vary along time. Video files can be considered a sequence of images and each image is called a frame [11]. Therefore, video files have more properties, as temporal resolution, which has a direct relationship with time and it is measured in frames per second (fps) [12].

One important characteristic is that neighboring frames have small variation. Thus, several techniques have been developed to perform video compression based on this characteristic [11]. Hence, the video frames are classified into two groups: independent frames and predicted frames. The Intracoded frames (I-frames) are complete images, encoded independently. The Predictive frames (P-frames) and Bidirectional frames (B-frames) have dependence on their neighbors, so they are not complete frames [12]. Therefore, these different frame types (and other existing in literature) can be used to compose the video stream.

B. Nudity and Pornography Detection in Videos

This subsection shows the main studies in the literature related to nudity and adult pornography detection in video files. These studies are important to illustrate the state-of-art of the subject, including the techniques used.

Poisel and Tjoa [9] presented a review of the state-of-art of forensic investigations of multimedia data. For automatic classification of video files, it highlights the usage of keyframes and motion analysis as the main techniques. Other techniques, like sliding window periodicity (PERWIN), the bag-of-visual-words (BOVW), periodicity detection (PER), and motion histograms are also cited for the detection of pornography in videos.

Pornography and nudity detection in videos through frame extraction was used by Kim et al. [13], Lopes et al. [10], and Wang et al. [14]. The algorithm developed by Kim et al. [13] also tries to detect global motions, separating the foreground objects, classifying the frames through skin color, shape, and texture features, comparing the objects with one built training database, using weighted Euclidean distance. Through techniques known as bag-of-visual-features (BOVF), previously developed for nudity detection in still images [15], Lopes et al. [10] used a gray-level SIFT and hueSIFT descriptors, generating a point descriptor that considers the color hue. Several images for training were used, building histograms. After that, a voting system classifies the video as nudity or not. Wang et al. [14] choose frames when changes in the video scene are detected. It tries to detect standards in video scene, such as fade in, fade out, dissolve, and wipe. The algorithms also try to identify human skin color, remove textures and unwanted image parts, before classifying the video as nudity or not. The technique takes about one tenth of the video duration to classify a video [14].

Unlike other initiatives of video frame extraction and subsequent detection in images, Endeshaw et al. [16], Jansohn et al. [17], and Rea et al. [18] used motion to determine pornography in videos. Endeshaw et al. [16] divide the video file in short fixed-length segments and create dominant motion vector for each frame. The algorithm tries to detect repetitive motion in a specific frequency band during 16 second intervals, classifying videos as decent or indecent. Jansohn et al. [17] calculate movement repetitions, called Periodicity Detection (PER), naturally involved in sex videos, and generate motion histograms. They combine frame static analysis, including bag-of-visual-words (BOVW), with the results of motion analysis. Rea et al. [18] used motion segmentation and built motion vectors to determine "homogeneous motion" in video files, which is used to determine if the video has pornography or not [18]. They also analyzed audio streams and measures audio periodicity. Zuo et al. [19] also used audio to recognize sounds of pornography. A framework was developed and it combines the detection results from video and audio to determine pornography in movies. The framework also uses extracted video frames to detect pornography through an image recognition algorithm.

As previously mentioned, the authors developed a study of automatic detection of child pornographic videos, based on frame extraction, with the definition of a Cutoff Point [8]. Such techniques are briefly discussed, including results achieved, in subsection C.3. This separation is necessary, because the proposed approach in this paper is an improvement of the prior work and it is directly related to the NuDetective Forensic Tool, reviewed below.

C. NuDetective Forensic Tool

This subsection provides a basic review of the main features of the NuDetective Forensic Tool [6, 7] to illustrate its basic operation, including its pixel analysis and human nudity detection algorithms in images. It also presents the previous techniques used to detect videos of child pornography [8].

1) Basic Features

The NuDetective Forensic Tool was developed to help forensic examiners to identify files of child pornography at crime scenes. To develop NuDetective, the authors considered two main principles: quick processing and reduced number of false negatives.

NuDetective Forensic Tool was developed with the Java Standard Edition (JSE), and has four main features (*Image Analysis*, *FileName Analysis*, *Hash Analysis*, and *Video Analysis*) and many other options that can be configured by the forensic examiner to assist in the detection of child pornographic files.

The usage of NuDetective is very simple: the forensic examiner chooses *locals* to search, such as folders and/or entire disk partitions, configures search options (optional task), and starts the scan for child pornographic files. Thus, the NuDetective scans all files of selected locals searching for files, submitting them to the internal detection algorithms, which can: detect nudity in images (*Image Analysis*); check the file names for common phrases of pedophilia (*FileName Analysis*); calculate the hash value of suspect files and compare them with a list of hash values (*Hash Analysis*); and also identify suspect videos of child pornography (*Video Analysis*).

The Tool was implemented using Java threads, allowing parallel execution of some tasks. One of the developed threads is responsible for searching the selected *locals*, the second cares only about the file analysis with internal algorithms, and a third thread is responsible for displaying the results, feeding the results GUI in real time. So the forensic examiner can start analyzing the results even before the Tool finishes searching all files of the selected *locals*.

2) Internal Nudity Detection Algorithms

The NuDetective Forensic Tool can automatically detect nudity in images [6]. It uses the RGB color space [20] for representing digital color space, which is the most common color space representation used for storing digital image [21]. To detect human skin colors in images, NuDetective classifies each image pixel as skin color or not, using a method [22] for detecting human skin color on RBG, as shown in the relation (1).

$$R > 95 \text{ and } G > 40 \text{ and } B > 20 \text{ and } max\{R,G,B\} - min\{R,G,B\} > 15 \text{ and } |R-G| > 15 \text{ and } R > G \text{ and } R > B \quad (1)$$

The Tool uses the algorithm presented by Apapid [23] for nudity detection, which does not require one previous database to classify the images, because the algorithm works with percentage of skin color and computational geometry.

Several experiments were performed to verify the robustness of the internal algorithms provided by the Tool. Previous results showed great automatic nudity detection rates, low false negatives, and quick processing [6]. The obtained values of *recall* and *precision* for nudity detection in images were 95.4% and 93.0%, respectively.

An optimization technique to reduce the runtime of automatic nudity detection in high-resolution digital images was also developed [7]. The optimization method reduces the runtime in high-resolution images in almost 90%, without significant variations in the results of nudity detection.

In literature, there are several other approaches for the detection of nudity and pornography in images using different color spaces and techniques [24, 25, 26].

3) Child Pornography Detection in Videos

In literature, as describe in subsection 2.B, it is possible to find many studies related to pornography detection in video, using a lot of different computational methods. All of them try to get the higher detection rate as possible and need a lot of processing time. So, besides being great methods to determine pornography in videos, they were not developed to be used at crime scenes. Also, the focus of these studies is not the specific detection of child pornography, but videos of nudity or pornography.

The *Video Analysis* feature provided by NuDetective needs to quickly identify videos of child pornography. Thus, in a previous study [8], a strategy was proposed. First, with 181 videos of child pornography, several experiments were carried out to understand the behavior and the nudity distribution in extracted frames along video duration, using the nudity detection algorithms implemented by the Tool. Later, the authors obtained the mean of nudity frames for each video and also the mean of nudity frames of all 181 video means ($\bar{X}$). The standard error (*E*) was calculated, and a Cutoff Point (*CP*) of 65.3% was determined, as shown in Figure 1. The Cutoff Point (*CP*) consists of a number that indicates the minimum percentage of nudity frames in a video to be classified as suspicious of containing child pornography.

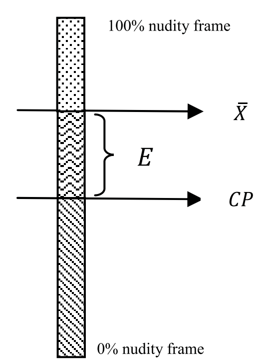

Figure 1: illustrates the mean of means $\bar{X}$, the standard error E, and the Cutoff Point *CP* [8].

After the definition of the Cutoff Point, several experiments were performed to define the amount of frames to be sampled at crime scenes, since processing all frames of a video requires a long time, being unfeasible at crime scenes. Two new sets of videos, one with 70 videos of child pornography and other with 79 common videos (not pornography), were used to define the sample size. The results showed that 10 frames uniformly sampled along each video is the optimum value to achieve low runtime, with good detection rates. Using that configuration, the experiments measured an average detection rate of 85.7%, with only 15.1% of false negatives, with values of *precision* and *recall* of 84.9% and 85.7%, respectively [8]. The developed algorithm could analyze all the 149 videos in only 308 seconds, i.e., an average of about only 2.06 seconds per video, including I/O time.

Therefore, the current strategy of sampling frames does not consider the video duration, always extracting exactly 10 frames for each video to be analyzed at crime scene.

D. Mathematical Functions

The mathematical functions have great importance in the modeling of various real life situations. A function can be defined by two variables: x (domain) and y (range). Each value of x corresponds to exactly one value of y, according to a specific definition [27]. The domain (x) of a function can be restricted according to the context in which the function is applied [28]. For example, the function x^2, used to calculate the area of a square, where x represents the square side, the domain should be restricted to positive numbers only.

An illustration of how different functions influence the values of y and x can be seen in Figure 2. The asymptotic behavior of the functions are substantially different: for smaller values of x, the functions I and II reach high values of y; the behavior is different in the functions IV and V, where the values of y grow more slowly when x is increased. The function V is a logarithmic equation and one of its main characteristic is to grow y slowly as x increases, tending to stabilize at infinity.

The general logarithmic function can be expressed as follows, in equation (2) [28]:

$$y = a + b\log_c(x - d), \tag{2}$$

Where a, b and c are constants, with $b \neq 0, c > 0$, $c \neq 1$, a and d are vertical and horizontal translations, respectively, b is vertical compression or vertical expansion, and c is the logarithm base.

Figure 3 illustrates graphical representations of logarithmic functions, showing how the variations of a, b and d have influence in their behavior.

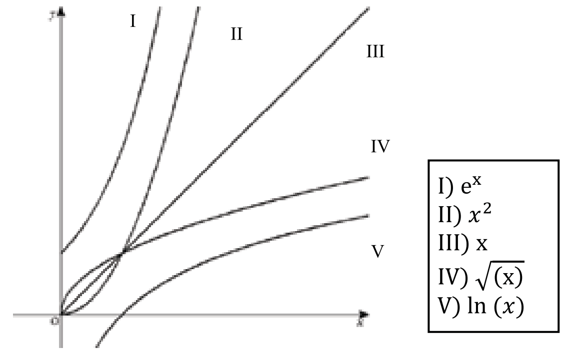

Figure 2: illustrates the behavior of different mathematical functions.

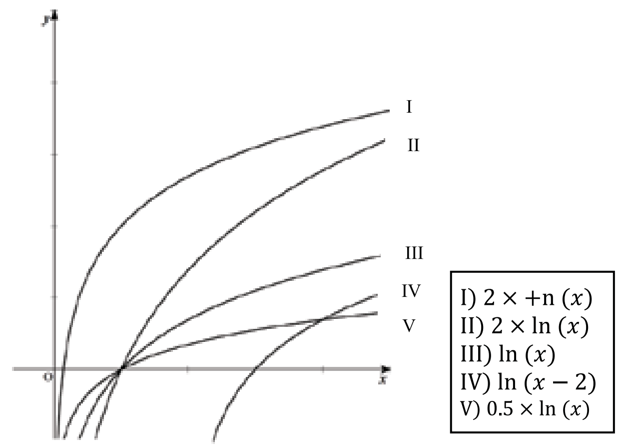

Figure 3: illustrates the influence of different values of *a*, *b*, and *d* of the equation (2).

III. Proposed Approach

This section presents the proposed adaptive approach of sampling video frames for the detection of child and teen pornographic videos at crime scenes. The adaptation consists in considering the size (in number of frames or duration) of the video to determine the amount of frames to be sampled. As described in Section II.C.3, a previous work [8] used a fixed number of samples in the detection of child pornography videos, achieving high detection rates and low runtime.

This work proposes the usage of a mathematical function to determine the number of frames to be sampled, considering the video duration, and trying to improve the detection rates and the runtimes previously achieved.

Considering the exposed context, the new adaptive approach must be quick enough to be used at crime scenes. Therefore, the number of frames to be sample must follow a mathematical function with specific behavior: it cannot increase y (number of samples) even in long movies or in videos with millions of frames. Thus, the use of an asymptotic function which tends to stabilize y is desired. As described in Section II.D, the logarithmic function has this behavior.

Thus, the logarithmic function was chosen to determine the number of frames to be sampled, because it can map the following assumptions:

1. The number of samples must vary with the video size/duration;
2. The number of samples cannot grow in the same proportion of the video size/duration;
3. Videos with great amount of frames or too long should not reflect significantly in the analysis time.

Using equation (2), it is possible to assign different values to the four constants in order to change the function behavior to determine, experimentally, the optimum setting for our objective. The constants a and d of equation (2) will not be used in our experiments, because they are responsible only for the translation of the graph and not by the change in the function asymptotic behavior. Therefore, discarding the constants a and d, it is possible to simplify the equation (2), obtaining the equation (3) below:

$$y = b \log_c x \tag{3}$$

The domain of this function, represented by the x, can be understood as the number of video frames or as its duration. The y represents the number of frames to be sampled. For example, with b equal to 1, c equal to 2, and x equal to 1024 frames, the number of frames to be sampled is 10, because:

$$10 = 1 \log_2 1024$$

In some cases, it may be important to use time (video duration) rather than total video frames as the population to be sampled. Although statistical sampling based on total number of video frames is correct, videos have a temporal resolution, which is important for defining the quality of the transition scenes and movements in the videos. However, temporal resolution is not important for the nudity detection algorithms implemented by NuDetective [6], because they are applied to static images (video frames).

To illustrate this problem, imagine one video scene with duration of 600 seconds. The scene was saved into two different video files with different temporal resolutions: the video 1 has 30 fps and the video 2 has only 1 fps. If the sampling strategy was based on the number of frames, considering the equation (3), b equal to 5 and c equal to 2, the total number of frames and the sample size of each video would be:

- Video 1: a total of 18,000 frames. Sample size of approximately 70 frames;
- Video 2: a total of 600 frames. Sample size of approximately 46 frames.

Thus, there is a clear distortion in the results, because the number of samples for the same scene is too different in the two videos, only because the variation of temporal resolution. The difference would be even greater if the sample was based on a fixed percentage, e.g., 1% of the frames (180 versus 6).

Whereas the two videos have the same scene, ideally the number of frames sampled should be equal. To correct this distortion, the equation (3) can be adapted to the equation (4) below, which is specific to each video. The equation (4) divides the total frames of the video (x) by its temporal resolution (*fps*), being adaptive to the duration (in seconds) of the video that will be analyzed. The number of frames to be sampled (y) should be rounded up.

$$y = \left\lceil b \log_c\left(\frac{x}{fps}\right) \right\rceil \tag{4}$$

The variables x and *fps* of equation (4) are specific to each video and they can be easy obtained at processing phase. However, to define the new sampling formula, is still required to determine ideal values for constants b and c. Using the same sets of videos previously used in [8], several experiments were performed by varying these two constants, obtaining, for each video, the number of frames to be sampled. Therefore, new detection rates and new runtimes for each value of these constants will be obtained, to empirically determine the optimal values for constants b and c, finally reaching the desired new adaptive formula.

Such experiments, including a real forensic case study, and the results are detailed in the next section.

IV. experiments and results

This section is divided into two parts: experiments to establish the best values for the equation constants (controlled experiments), and validation of the proposed approach in a real forensic case experiment (validation experiment).

A. Empirical establishment of equation constants

This subsection describes the experiments that were conducted to determine empirically the best values for the constants b and c (base) of the equation (4), responsible for defining the number of frames to be sampled from a given video.

The results were evaluated according to the values obtained for detection rate and runtime. For the experiments, the authors used the same hardware as in [8] (laptop with 2.33 GHz Intel Centrino Duo processor, 2GB RAM, and running Windows 7), and the same video database. The goal is to allow a direct comparison of this work with the previous one [8], on the two aspects (detection rates and runtimes). The video databases used for the experiments have the following characteristics:

- Test Set 1: 70 videos of child pornography obtained from seized hard disks of real forensic cases;
- Test Set 2: 79 common videos, such as interviews, sports, movie trailers, music video clips, among others not related to pornography, obtained from the Internet.

To determine the influence of the constants b and c of the logarithmic function, described in the equation (4), were used b values of 1, 2, 3, 4, and 8, and c values of 2 and 10. Thus, for each video of test sets, the number of frames to be sampled was calculated. Frame samples were extracted from positions uniformly distributed throughout the video. All frames were analyzed by the nudity detection algorithm provided by NuDetective, as described in section II.B.2, and the video is classified as suspicious of containing child pornography if the percentage of nudity frames is greater than or equal to the Cutoff Point (65.3%). In some cases, the *CP* was rounded to the lowest possible value of sampled frames. The Tables 1 to 4 show the results of these experiments.

Table 1: results obtained from several values of *b*, and <u>*c* equal to 2</u>, in *Test Set 1* (70 videos of child pornography).

b	Classified as child pornography	Success rate	Runtime (sec.)
1	57	81,4%	55
2	58	82,9%	113
3	56	80,0%	182
4	54	77,1%	226
8	54	77,1%	503

Table 2: results obtained from several values of *b*, and <u>*c* equal to 10</u> in *Test Set 1* (70 videos of child pornography).

b	Classified as child pornography	Success rate	Runtime (sec.)
1	62	88,6%	16
2	61	87,1%	32
3	61	87,1%	48
4	59	84,3%	66
8	56	80,0%	135

Table 3: results obtained from several values of *b*, and c equal to 2 in *Test Set 2* (79 videos not related to pornography).

b	Classified as child pornography	Success rate	Runtime (sec.)
1	9	88,6%	124
2	8	89,9%	253
3	10	87,3%	415
4	9	88,6%	514
8	10	87,3%	1127

Table 4: results obtained from several values of *b*, and c equal to 10 in *Test Set 2* (79 videos not related to pornography).

b	Classified as child pornography	Success rate	Runtime (sec.)
1	17	78,5%	35
2	13	83,5%	69
3	10	87,3%	122
4	12	84,8%	146
8	9	88,6%	297

To determine which values of b and c achieved the best results, it is important to analyze the results of detection rates and runtime in both test sets. The results obtained (Tables 1 to 4) show the increase in runtime as the value of b increases, which is expected, since b is directly proportional to the number of samples. However, detection rates did not improve when the sample size increased, rather, in most cases, decreased.

Analyzing the results, it is possible to identify that the values b equal to 3 and c equal to 10, achieved the best results of accuracy and runtime, considering the two test sets together. For these parameters, the values of *recall*, *precision,* and runtime are shown in Table 5, together with the previous obtained values [8], when all the videos had 10 frames extracted as samples. The equation (5) shows the best configuration of constants.

$$y = \left\lceil 3\log_{10}\left(\frac{x}{fps}\right)\right\rceil \quad (5)$$

In equation (5), x is the total number of frames of the video, *fps* the temporal resolution of the video, and *y* the rounded up number of frames to be sampled for the video. The Table 5 also shows the values of *F-measure* [29], which combines the values of *precision* and *recall* in a unique value. The equation (6) was used for this calculation with which indicates the same weight for *precision* and *recall.*

$$F_\beta = \frac{(\beta^2+1) \times \text{recall} \times \text{precision}}{(\beta^2 \times precision)+recall} \quad (6)$$

Table 5: comparison between the results of the proposed approach in this work versus previous results [8].

Measure	Previous Experiment	Current Experiment	Difference
Recall	84.9%	87.3%	+2.4%
Precision	85.7%	85.9%	+0.2%
F-measure	85.3%	86.6%	+1.3%
Runtime	308 seconds	170 seconds	-44.8%

The results showed that the new adaptive strategy has a small detection rate improvement, but the runtime has dropped significantly, being around 45% faster than the previous strategy. This was only possible because the proposed sampling approach adapts to each video: short videos have smaller samples, and longer videos have larger samples, optimizing the sampling procedure according to the characteristics of the videos to be analyzed at the crime scenes.

B. Validation of proposed approach in a real forensic case

This subsection shows the performed experiment to validate the proposed approach. The new adaptive strategy of child pornography detection in digital videos was implemented in the *Video Analysis* feature provided by NuDetective Forensic Tool.

For the experiment, we used a 160 GB hard disk drive (HDD) seized during an operation to combat pedophilia in Brazil. The HDD was placed in a *Forensic Logicube Dossie* that was connected to the USB port of the same laptop computer previously used. This configuration, which is widely used by forensic examiners at crime scenes, ensured the preservation of all data in the HDD, because the *Dossie* acted as a write blocker of the storage device.

Using the laptop, the NuDetective searched the entire disk for videos of child pornography. Only the feature *Video Analysis* of NuDetective was used with default options – the other options were disabled (*Image Analysis*, *FileName Analysis* and *Hash Analysis*). To compare the results achieved with NuDetective, the HDD was also analyzed in a forensic laboratory, using commercial forensic tools, such *AccessData Forensic ToolKit* (FTK) and *Guidance EnCase*, to determine the real amount of videos containing human nudity and child pornography stored in the HDD.

The NuDetective took only 137 seconds to search the entire content of the HDD for video files. A total of 87,576 files were stored in the HDD, with 401 videos of various types, lengths and sizes. The NuDetective automatically identified and processed all the 401 videos and classified 186 video files as suspicious, showing them in the results GUI. Among the 186 videos identified as suspicious, 175 were in fact related to nudity, including child pornography. As expected, the tool significantly reduced the amount of files to be examined by the forensic examiner, from 87,576 saved files on the HDD to only 186 suspicious files. Table 6 shows the confusion matrix [30] obtained from this validation experiment in a real forensic case. The obtained values of *precision* and *recall* in the validation experiment were 93.0% and 80.0%, respectively, and the *F-measure* was 86.0%.

Table 6: Confusion matrix sumarizing the results of the validation experiment (real forensic case).

		True Class	
		Positive	Negative
Classified as	Pos.	79.5%	6.0%
	Neg.	20.5%	94.0%

The results of the new video detection strategy identified approximately 80% of videos of child pornography stored in the seized HDD in only 137 seconds. Therefore, it was proved that this new adaptive strategy for detecting videos of child pornography can be used to identify these illegal files, filtering out the files to be analyzed by the forensic examiner at crime scenes.

V. Conclusions

This paper proposed an improvement of the frame sampling strategy to identify child pornographic videos at crime scenes. The previous strategy used 10 frames of each video to be automatically analyzed, which, despite the great detection results and low runtime, did not consider the video duration to determine the frame sample size. Thus, this work carried out a series of experiments with the same 149 videos originally tested, which 70 are videos of child pornography, to develop a new formula to sample frames.

To achieve the main objective, the logarithmic function was chosen due to its characteristic of nearly stabilize *y* values at infinity, ideal for long videos, because it is not feasible to analyze large amount of frames at crime scenes. By varying the multiplier coefficient and the base, the logarithmic experiments showed different detection rates and runtimes. After analysis results, the best configuration obtained were multiplier coefficient 3 and base 10, as shown in equation (5).

The new proposed formula achieved detection rates around 87% and runtime of only 170 seconds, i.e., an average of 1.14 seconds per video, including I/O time. The new strategy reduced the runtime in 45% and achieved a small improvement in the detection rates. The new values for *precision* and *recall* were 85.7% and 87.3%, respectively.

In a real forensic case of pedophilia, the new strategy proved to be a great way to identify child pornographic videos, detecting around 80% of these illegal videos saved on a 160 GB hard disk drive. Therefore, the NuDetective found and analyzed 401 videos in only 137 seconds, an average of approximately 0.35 seconds per video, including I/O time, being much quicker than in controlled experiments. The obtained *F-measure* values were 86.6% and 86.0% in the controlled and validation experiments, respectively.

The results achieved great reduction in runtime due to adaptation of the strategy proposed to consider the video duration, making shorter videos to be processed quickly, i.e., with less number of sampled frames without compromising detection rates. However, for longer videos, the strategy used more samples without compromising the runtime. Thus, since child pornographic videos are generally short, this new strategy reduced around 45% of the runtime to identify these files, which is desired at a crime scenes. As mentioned before, the achieved results of runtimes are directly related to the equipment used and, in all these conducted experiments, we used a technologically outdated laptop. So, using new computers, the runtime may be even lower.

It is noteworthy that the previously calculated Cutoff Point of 65.3% was maintained. The new adaptive strategy was implemented in the *Video Analysis* feature provided by NuDetective Forensic Tool. Thus, the new version of the tool has a more effective and robust detection strategy of child pornography videos, helping even more the forensic examiners to identify such files quickly and reliably at crime scenes.

The NuDetective Forensic Tool is free and available for law enforcement use only. For instructions on how to obtain the Tool, please contact the authors or send an e-mail to nudetective@gmail.com.

Acknowledgment

The authors thank to the Brazilian Federal Police for financial and logistic support.

References

[1] Baines, V. Online Child Sexual Abuse: The Law Enforcement Response. A contribution of ECPAT International to the World Congress III against the Sexual Exploitation of Children and Adolescents, Rio de Janeiro, Brazil, 2008.

[2] Saroiu, S. et. al. A Measurement Study of Peer-to-Peer File Sharing Systems. Proceedings of the Multimedia Computing and Networking (MMCN). San Jose, Jan, 2002.

[3] Choo, K-K. R. Online child grooming: a literature review on the misuse of social networking sites for grooming children for sexual offences. Australian Institute of Criminology, pp 48-66. Available from http://www.aic.gov.au/publications/current%20series/rpp/100-120/rpp103.aspx; [Visited Jul, 2012].

[4] BRAZIL. Law 11,829, November 25th, 2008. Brazil, 2008; [in portuguese]

[5] Eleuterio, P. M. S.; Machado, M. P. Identificação de autoria e materialidade em crimes de abuso sexual de criança/adolescente a partir da análise de arquivos multimídia. IV International Conference On Forensic Computer Science (ICoFCS 2009), Natal/RN, Brazil, 2009; [In portuguese].

[6] Polastro, M. C.; Eleuterio, P. M. S. NuDetective: a Forensic Tool to Help Combat Child Pornography through Automatic Nudity Detection. Proceedings of the Twenty-First Internation Workshop on Database and Expert Systems Applications (DEXA'10 – VISM'10) - pgs. 349-353. IEEE Computer Society. Bilbao, Spain. Aug, 2010.

[7] Eleuterio, P. M. S.; Polastro, M. C. Identification of High-Resolution Images of Child and Adolescent Pornography at Crime Scenes. THE INTERNATIONAL JOURNAL OF FORENSIC COMPUTER SCIENCE - IJoFCS, v. 5, p. 49-59, 2010.

[8] Polastro, M.C.; Eleuterio, P.M.S. A statistical approach for identifying videos of child pornography at crime scenes. Proceedings of the Fifth Workshop in Digital Forensics (WSDF'12 – ARES'12). IEEE Computer Society. Prague, Czech Republic. [To apeear in Aug, 2012].

[9] Poisel, R.; and Tjoa, S. Forensics investigations of multimedia data: A review of the state-of-the-art. Proc. of the 6th International Conference on IT Security Incident Management & IT Forensics (IMF'11), Stuttgart, Germany. IEEE, May 2011, pp. 48–61.

[10] Lopes, A.P.B., Avila, S.E.F., Peixoto, A.N.A., Oliveira, R.S., Coelho, M. de M. & Araújo, A. de A. Nude detection in video using bag-of-visual-features. Proceedings of the XXII Brazilian Symposium on Computer Graphics and Image Processing, SIBGRAPI, IEEE Computer Society Press, Rio de Janeiro, RJ, Brazil, 2009, pp 224-231.

[11] Halsall, F. Multimedia Communications: Applications, Networks, Protocols, and Standards. Addison-Wesley Publishing, 2000.

[12] Steinmetz, R.; Nahrstedt, K. Multimedia Fundamentals, Volume I: Media Coding and Content Processing (2nd Edition). Prentice Hall, 2002.

[13] C.-Y. Kim, O.-J. Kwon, W.-G. Kim, and S.-R. Choi. Automatic System for Filtering Obscene Video. In ICACT, volume 2, pages 1435--1438, 2008.

[14] Wang, D.; Zhu, M.; Yuan, X.; Qian, H. Identification and annotation of erotic film based on content analysis. Proc. SPIE 5637, 88 (2005).

[15] Lopes, A.P.B., Avila, S.E.F., Peixoto, A.N.A., Oliveira, R.S. & Araújo, A. de A. A bag-of-features approach based on Hue-SIFT descriptor for nude detection (in press). Proceedings of the 17th European Signal Processing Conference, EUSIPCO, Glasgow, Scotland, 2009, pp. 1552-1556.

[16] Endeshaw, T.; Garcia, J.; Jakobsson, A. 2008. Classification of indecent videos by low complexity repetitive motion detection. In *Proceedings of the 2008 37th IEEE Applied Imagery Pattern Recognition Workshop* (AIPR '08). IEEE Computer Society, Washington, DC, USA, 1-7.

[17] Jansohn, C.; Ulges, A. & Breuel, T. M. Gao, W.; Rui, Y.; Hanjalic, A.; Xu, C.; Steinbach, E. G.; El-Saddik, A. & Zhou, M. X. *(ed.)* Detecting pornographic video content by combining image features with motion information. *ACM,* 2009, 601-604.

[18] Rea, N.; Lacey, G.; Lambe, C.; Dahyot, R. Multimodal Periodicity Analysis for Illicit Content Detection in Videos. *Visual Media Production, 2006. CVMP 2006. 3rd European Conference on* In Visual Media Production, 2006. CVMP 2006. 3rd European Conference on (2006), pp. 106-114.

[19] Zuo, H.; Wu, O.; Hu, W.; Xu, B. Recognition of blue movies by fusion of audio and video. Proceedings of the IEEE International Conference on Multimedia and Expo(ICME), pp. 37–40, April 2008.
[20] Vezhnevets, V.; Sazonov, V.; Andreeva, A. A Survey on Pixel-Based Skin Color Detection Techniques. Proceedings Graphicon, 2003.
[21] Kakumanu, P.; Makrogiannis, P.; Bourbakis, N. A survey of skin-color modeling and detection methods. Pattern Recognition 40, 2007.
[22] Kovac, J.; Peer, P.; Solina, F. Human skin colour clustering for face detection. International Conference on Computer as a Tool EUROCON 2003, Eds. B. Zajc, Ljubljana, Slovenia, September 2003.
[23] Ap-apid, R. An Algorithm for Nudity Detection. Proceedings of the 5th Philippine Computing Science Congress, Cebu City, 2005.
[24] Carlsson, A.; Eriksson, A.; Isik, M. Automatic detection of images containing nudity. Master thesis in intelligent systems Design, University of Gothenburg, Sweden, 2008.
[25] Lin, Y.; Tseng, H.; Fuh, C. Pornography Detection Using Support Vector Machine. 16th IPPR Conference on Computer Vision, Graphics and Image Processing (CVGIP). Kinmen, ROC, 2003.
[26] Wang, J.; Li, J.; Wiederhold, G.; Firschein, O. System for Screening Objectionable Images Using Daubechies' Wavelets and Color Histograms. Proc. of the International Workshop on Interactive Distributed Multimedia Systems and Telecommunication Services, pages 10–30, 1997.
[27] Hwang, A. D. Calculus for Mathematicians, Computer Scientists, and Physicists - An Introduction to Abstract Mathematics. Holy Cross, 1998.
[28] Foerster, P. A. Precalculus With Trigonometry: Concepts and Connections. Key Curriculum Press, Emeryville, CA, USA, 2003.
[29] Van Rijsbergen, C. J. Information Retrieval. Butterworths. 1979.
[30] Fawcett, T. An introduction to ROC analysis. Pattern Recognition Letters 27 (2006) 861–874.
[31] A. Vrubel, "Creation and Maintenance of MD5 Hash Libraries, and their Application in Cases of Child Pornography", The International Conference on Forensic Computer Science (ICoFCS), 2011. Available: http://dx.doi.org/10.5769/C2011015

Mateus de Castro Polastro obtained his bachelor's degree in Computer Science at the University of Campinas (Unicamp), Brazil. He obtained his Master's degree in Computer Forensics at the University of Brasilia (UnB), Brazil. Since 2007, he has worked as a criminal forensic expert for the Brazilian Federal Police (DPF).

Pedro Monteiro da Silva Eleuterio is a Computer Engineer at the Federal University of Sao Carlos (UFSCar), Brazil. He obtained his Master's degree in Computer Science at the University of Sao Paulo (USP), Brazil. Since 2006, he has worked as a criminal forensic expert for the Brazilian Federal Police (DPF). He is also the author of the Computer Forensics book "Desvendando a Computação Forense" (in Portuguese).