UNIVERSITATEA POLITEHNICA BUCUREȘTI

FACULTATEA DE AUTOMATICĂ ȘI CALCULATOARE

DEPARTAMENTUL CALCULATOARE

|  |  |
| --- | --- |
| upb | cs |

PROIECT DE DIPLOMĂ

Detecția imaginilor cu impact emoțional

și clasificarea tipurilor de violență

Edwin-Mark Grigore

**Coordonator științific:**

Conf. dr. ing. Traian Rebedea

BUCUREŞTI

2020

UNIVERSITY POLITEHNICA OF BUCHAREST

FACULTY OF AUTOMATIC CONTROL AND COMPUTERS

COMPUTER SCIENCE DEPARTMENT

|  |  |
| --- | --- |
| upb | cs |

DIPLOMA PROJECT

Graphic Content Detection in Images and

Classification of Types of Violence

Edwin-Mark Grigore

**Thesis advisor:**

Lect. Dr. Engr. Traian Rebedea

BUCHAREST

2020

**CONTENT**

[Sinopsis 2](#_Toc42265991)

[Abstract 2](#_Toc42265992)

[1 Introduction 3](#_Toc42265993)

[1.1 Background 3](#_Toc42265994)

[1.1.1 Children exposure to violence 3](#_Toc42265995)

[1.1.2 Parental Control Applications 3](#_Toc42265996)

[1.1.3 Classification of violence and graphic content 4](#_Toc42265997)

[1.2 Problem Description 5](#_Toc42265998)

[1.3 Goals 6](#_Toc42265999)

[1.3.1 Dataset 6](#_Toc42266000)

[1.3.2 Augmentation 7](#_Toc42266001)

[1.3.3 Training the model 7](#_Toc42266002)

[1.4 Outline of the Thesis 8](#_Toc42266003)

[2 Project specifications 9](#_Toc42266004)

[3 State of the Art 10](#_Toc42266005)

[3.1 Violent Images Dataset 11](#_Toc42266006)

[3.1.1 Violent Scene Dataset by InterDigital 11](#_Toc42266007)

[References 13](#_Toc42266008)

# 

# Sinopsis

Copiii sunt cel mai ușor de influențat. Fie că acest lucru se face prin cuvinte sau prin fapte, prin ceea ce văd sau prin ceea ce aud, ei sunt la vârsta la care cel mai mic lucru își poate lăsa amprenta asupra lor. Ceea ce privesc, imaginile și filmele pe care le vizionează, chiar și știrile sau noutățile de care pot auzi și cu care intră în contact, pot avea un impact asupra lor. O imagine nepotrivită ajunsă sub ochii unui copil de 5 ani îl poate traumatiza pe viață sau, cel puțin, îi poate genera o temere de care să nu poată scăpa niciodată.

Tehnologia a evoluat, s-a extins într-atât de mult, încât este prezentă în viețile tuturor în cele mai mici detalii, și din ce în ce mai pregnant, la îndemâna copiilor. Pornind de la această realitate și pentru a îmbunătăți experiența aplicațiilor de parental control, este necesară detecția imaginilor ce conțin scene de violență sau scene cu impact emoținal puternic, imagini ce conțin sânge sau înfățisează corpuri umane având plăgi deschise sau cadavre, incendii violente sau prezența armelor.

# Abstract

The children are the easiest to influence. Whether this is done by words or by actions, through what they see or what they hear, they find themselves at the age when the smallest thing can leave a mark on them. Anything they watch, the images, the movies, or even the news they hear or happen to come across, can have a major impact on their lives. An inappropriate image that comes under the attention of a 5-year-old child, can traumatize him for life, or at least cause him severe fear that he would never be able to get rid of.

The technology has evolved and expanded so much over the last decades, that is present in everybody’s life in every little aspect, and more and more significantly at children’s disposal. Starting from this reality and for enhancing the experience of parental control applications, it is necessary the identification of the images that contain scenes of violence or emotionally disturbing scenes, images that contain blood or depicts human bodies with open wounds, violent fires or presence of guns and weapons.

# Introduction

## Background

### Children exposure to violence

The digital era has allowed most people of the planet to be connected to the Internet via continuously shrinking devices: personal computers, smartphones, smart gadgets, smartwatches etc., meaning anyone has access to any kind of information at any moment. Alongside the obvious and huge advantages Internet, media and technology altogether have brought to the world, there are many risks worth noting and addressing.

Focusing on the rapid spread of smart gadgets in the homes of most of the people, forth comes an issue that has been long debated and argued: children exposure to violence through media. Whether is television or movies, games, news, everything a child hears and perceives has an impact on him/her. It is in the hands of the parents to prevent their children from seeing images that would cause them anxiety, fear, misunderstanding of society, or even traumas.

The common adage “A picture is worth a thousand words” denotes exactly how an image can influence a child, especially if we are talking about inappropriate images. Studies [1]have shown that aggressive or antisocial behaviour is heightened in children after watching violent television or films. Science tells us that early exposure to extremely fearful events affects the developing brain, particularly in those areas involved in emotions and learning [2]. When children see images that are emotionally disturbing, images that depicts the world in an inadequate manner for their young minds to comprehend, they can learn fear from situations they should not be exposed to. The brain stores what we see, and how often do we hear people say, “I have seen something I will not forget my entire life”. If that is the case for mature persons, then even more importance should be given to what children are allowed to see.

### Parental Control Applications

Most of the online media platforms that are available to children and have no age restrictions by default – i.e., film-producer or film-provider platforms, video-sharing platforms, applications stores, online electronic bookstores etc. – offer mechanisms of protection targeted on age, mainly named *Parental Control*, that will allow the parents to set restrictions on the content their children must not access.

Recently, there have been developed several parental control applications by different companies in order to aid parents overcome the issues the ever-growing technology comes along with. Many give the possibility to the parents to track their children’s both online activity (i.e. accessed websites) and offline activity (e.g. applications used, time spent on devices etc.), as well as to keep a record of the people their children interact with via calls or chats.

The potential dangers are identified through complex artificial intelligence algorithms that would analyse and detect any kind of inappropriate content on their devices, whether it is an image sent in a chat or a bullying text message received, subsequently alerting the parents about the problem and blocking the threat.

This way, the children are protected from being exposed to any harmful content or situation they may encounter while browsing the Internet or chatting with others.

### Classification of violence and graphic content

The term *violence* is “used to animal and human behaviour that threatens to cause or causes severe harm to a target.” [3]

*Violence* is “the intentional use of physical force or power, threatened or actual, against oneself, another person, or against a group or community, which either results in or has a high likelihood of resulting in injury, death, psychological harm, maldevelopment, or deprivation.” [4]

There has been much controversy about this term and about what should be included under the heading of violence. The World Health Organization divided violence into three categories: self-directed violence, interpersonal violence and collective violence [4]. These three categories are divided further to depict more specific types of violence: physical, sexual, psychological, emotional.

Transposing the notions to the field of images, the violence transcends these categories. The term *graphic* or *graphic violence* refer to depiction acts of violence in visual media such as film, television and video games. The violence may be real or simulated.

Graphic violence generally consists of uncensored depiction of various violent acts which includes depiction of murder, assault with a deadly weapon, accidents which result in death or severe injury, and torture.

Graphic violence causes intense emotions, especially in children, and frequent exposure can lead to desensitization over the severity of the actions depicted. Over time, children may become too familiar with violence and may learn that a little violence is socially acceptable, mainly because it is present in many contexts, such as films, news, literature, that children are accustomed to and take as a natural thing.

Many online media platforms have defined standards and policies regarding this subject and some have restricted, or even prohibited inappropriate content on their platforms. Social-media platforms (e.g. Facebook) have introduced the guidelines for their users, which act as rules to be followed when interacting with other users.

Facebook wrote the *Community Standards* [5] which stipulates the rules that platform users should take into consideration when posting. On this basis, they promote and fight to maintain a “safe environment on Facebook” by removing any content that is identified as graphic, upsetting or harmful for viewers, depicting different kinds of violence in images, videos or speech. Facebook divides prohibited content into several categories, starting with hate speech, listed as a direct attack based on race, ethnicity, religious affiliation, sexual orientation, gender or disability and all the way to nudity and sexual activity, violence and criminal behaviour, which include images or videos of people or dead bodies in non-medical settings, child or adult abuse, live streams of capital punishment of a person, still images depicting the violent death of a person. Facebook have their own AI algorithms that detects such images and videos and moderators that apply the rules, by censoring and removing posts that do not follow the rules.

In this context, this project would like to address this issue with technology in respect of its impact on children and their healthy emotional and psychological growth.

## Problem Description

In order to prevent the exposure of children to graphic and violent images, these images must be identified first. Since parents can not be physically near their children every single time, they must rely on the technology they use to do that.

As the technology advances, Computer Vision leads the way in training artificial intelligence to learn to interpret and understand the visual world. Using deep learning models, we now have machines that can accurately identify and classify objects. Computers can acquire and process data, analyse and understand the content of the images, and they can extract the information necessary to perform different tasks and decide based on the content of the image.

Although there is extensive knowledge to develop such deep learning models, only a few have been created that recognize and/or classify the violence depicted in still images, and even less are available to the public use.

There are several potential areas that may use machine learning to detect violence, such as parental control applications and web filtering, that makes this subject worth being studied.

The human brain can perceive an image and understand the content of the image in the blink of an eye. A team of neuroscientists from MIT found that the brain can process an image even if the eyes see it for only 13 milliseconds [6]. They used rapid serial visual presentation (RSVP) of series of 6 to 12 pictures presented at between 13 to 80 milliseconds per picture. The participants were asked to detect a picture specified by name, which they did. The accuracy of detection improved with increasing the duration of presentation, nevertheless the result of the study shows how fast humans can understand the meaning of a picture.

Judging an image whether it contains violence or not it is slightly more difficult. While most of the time, one can do it relatively fast, some situations take a little longer for an answer to be given.

Assigning this job to a computer is no easy task. Computer Vision includes several algorithms that will successfully identify objects in images with almost 100% accuracy. However, identifying the graphic content is more than simply object identification.

The depiction of violence is relative. The brain perception of two persons is different, so one may judge an image as being graphic, while the other may not. Some images that depict the exact same objects, but varying in pose and instance, may convey violence in some cases, but no violence at all in other cases, depending on the context in which those particular objects are depicted. An example, and maybe the most difficult to work with, is fire. An explosion can easily be classified as graphic content, but a fire camp might be mistakenly classified as being upsetting for viewers due to its size or presence of people around it.

The difficulty of the problem comes from this aspect. The context depicted in the images or the circumstances under which the photograph appear to have been taken may be misleading for the artificial intelligence, while for the human eye it would be an easy decision.

## Goals

This project has the following goals:

* Gather a dataset of images that depict the best graphic content and violence
* Augment the dataset
* Train a deep model to recognize and classify the types of violence in still images

### Dataset

The purpose of this stage is to gather as many useful sets of data that can be used to accurately train the model. The images must be labelled and must portray the type of violence labelled precisely, clearly. The dataset should not contain any images that are inaccurate.

The dataset can be acquired online downloading from a reliable source or by manually browsing the search-engines and downloading images. The goal is to find a dataset that has been already created, because that will save a lot of time. Ultimately and as a last resort, the second option should be taken into consideration.

After acquiring the dataset, we want to go through the directories and prune irrelevant images (e.g. images that can be double classified, over- or underexposed images where the subject is not visible enough). The trickiness of this step comes from the fact that multiple images can be classified both violent and non-violent by human intelligence, depending on who is making the call and based on the context. This kind of images must be carefully separated into the corresponding labels, so an image that doesn’t show a violent or graphic content, but is similar in composition to one which does, would not end up being mislabelled.

At the end of this step, the dataset should be ready for use and it should consist of as many images as possible. A solid start would be 1000 images per category.

### Augmentation

Given the fact that the dataset might be small, despite the efforts, the next step is to extend the size of the dataset. The augmented images are acquired by performing a series of transformation on the base images, that include rotation, cropping, skewing, zooming, flipping and mixing.

Because neural networks learn the features of the images, these features must be represented in as many ways as possible for the trained model to be able to recognize these features in images that it will be used on.

The augmentation can solve numerous problems of the dataset. It does not only increase the size of the dataset considerably, but it can also bring in diversity given by the subtle variation in orientation, position, angle, rotation etc. of the subject depicted. This will result in the dataset to contain, for example, a specific weapon in many positions on an image, increasing the accuracy of the model to be trained.

At the end of this step, the dataset size should increase several times and should depict a diversity of features.

### Training the model

The final goal of the project is the training of the model. At this moment we have a good dataset, which has been augmented and ready to use. This dataset must be used to train a model to recognize violence and graphic scenes in images with a high accuracy.

Transfer learning is used in order to achieve the goal. This method applies different existing models that have already been trained for general purposes, to the characteristics of this project.

Transfer learning can save a lot of training time, given the fact that we start from something that already exists and has already been trained on actual data. Most of the times, the machine learning model that is used has been trained on datasets much larger and complex than our dataset and for a longer time than we can do for our own model. A broad dataset is crucial in feature learning and this will be the case for this project.

There are several steps that must be taken in order to train the model. First, we must choose from the large pool of the existing deep learning models one to be the basis to our training. The process of identification of the model that works best on the dataset available is believed to be a key aspect. Second, we take the pre-trained model and use it as a starting point for our model. Also, we must decide which layers of the pre-trained model that we will use in the process and what layers be must build on top of it. Finally, we must adapt and refine the pair of so it may fit as well as possible the task at hand, process called tuning the model.

## Outline of the Thesis

This thesis presents a project with the goal of creating a machine learning model that can identify and classify violence and other graphic content in still images. The output model will be able to identify the followings: presence of weapons, presences of fire, fight scenes, presence of blood and gore, detailed in Chapter 2.

The Chapter 3 will discuss the existing machine learning models, the current technology and alternatives to this project. Then, Chapter 4 describes the implementation step by step, and finally, the results will be analysed, and conclusions will be drawn in Chapter 5.

# Project specifications

The project aims to provide a solution to identifying potential harming and upsetting images for children up to 14 years old.

The main features the output machine learning model of this project intends to identify and classify are presented in the following lines.

*Presence of firearms –* the presenceof any type of gun or similar fire weapon, regardless of the intent of the subject depicted, is to be classified as violent image.

*Presence of cold weapons –* any type of melee weapon, ranged weapon or other type of weapon that does not involve fire or combustion, is to be classified as violent image.

*Presence of fire* – any explosion, big fire, vegetation fire, human or animal, living or dead that is burning, fire caused by a gun, is to be classified as violent.

*Fight scenes –* any fight, regardless of the number of people involved or how they are fighting or the weapons they use, is to be classified as violent image.

*Presence of blood and gory scenes* – any serious body injury, any presence of blood that drains out from a body, any wound or tissue damage, presence of horror creatures, mutant creatures or skull and flesh representation, is to be classified as violent image.

Any other image that does not contain violent depictions or graphic contents, and that is not classified under the above categories, will be labelled as *non-violent*.

The model may be subsequently integrated into a parental control application, providing the means to detect if the images a user opens, receives or sends, are graphic or contain violence and notify the parents about the potential harmful situation.

# State of the Art

Deep learning has seen an immense increase in interest over the last few years. Since the term *deep learning* was introduced in 1986 by Rina Dechter [7] to the machine learning community, many steps have been taken to elaborate and build the architecture we know today. Deep learning has since been applied to fields including computer vision [8][9], speech and audio recognition, social network filtering, many producing results compared to or even surpassing the human ability [10].

It is called *deep* because the architecture uses multiple layers in the network. The unbounded number of layers allows an optimized implementation and has practical application.

The countless work that has been done in this field is great and listing all the research that have brought deep learning and computer vision to what it is today is beyond the purpose of this paper. One of them, however, is worth noting.

Believing that, despite all the hard work, research and discoveries made, something was still missing, the Stanford University professor, Fei-Fei Li launched in 2009 the ImageNet [11]. It was something that had not been done until that time. Fei-Fei Li thought that the missing part was the data. And while there were images all over the Internet, there was no such thing as a database containing labelled images. Therefore, she assembled a free database consisting of more than 14 million labelled images, saying, “Our vision was that Big Data would change the way machine learning works. Data drives learning.”[[1]](#footnote-1) They used the dataset in the annual Large-Scale Visual Recognition Challenge (LSVRC) where contestants would build computer vision models which were rated by their accuracy.

In 2012, Alex Krizhevsky, Ilya Sutskever, and Geoff Hinton submitted a model that was almost two times better than the previous ones, with an error rate of 15.3%. They used GPUs to train the model, as their power of computing offered a great amount of speed. They also reduced overfitting by introducing the method called *dropout* and used the *rectified linear activation unit* (ReLU), methods that later became the basics to deep learning. The network is known as *AlexNet.* [9]

Since then, computer vision and deep learning have seen a rise in interest, research and applications. Deep learning is now present almost in any domain of activity: defeating the world best player at Go [10], AI-driven cars [12], diagnosing skin cancer [13] or surveillance identification, only to name a few.

## Violent Images Dataset

Since ImageNet, several images datasets have been assembled and are available for free. They gather and label images for thousands of categories, each consisting of hundreds of images. The categories of these datasets are various, so the dataset can train a model more precisely in order to serve a broad purpose in visual recognition.

According to specific needs of different projects, many smaller datasets have been created, each consisting of images from different categories that depict the subject those models should recognise, which would allow them to train a network over an existing one that has already been trained with a huge dataset such as ImageNet. For this purpose, some datasets of guns and weapons images that have been used to train models that can detect weapons in images or videos can be found on the Internet[[2]](#footnote-2) and are free to use. [18]

However, the number of datasets of images containing fight scenes or gory scenes is low, and even those are very small and not concluding. A slightly more convincing number of datasets can be found for videos depicting violence, but still, when it comes to fights or blood presence, there are not too many.

### Violent Scene Dataset by InterDigital

Violent Scene Dataset (VSD) created by InterDigital is a public dataset for the detection of violent scenes in videos. It is a collection of labels, features and annotations based on the extraction of violent scenes from films and web videos. It also contains audio annotations of violence-depicting sounds present in the videos [14][16][17]. According to the description made in the publications [15], the data was produced by the Fudan University and the Ho Chi Minh University of Science.

The dataset consists of 86 short videos downloaded from YouTube and normalised to a frame rate of 25. Also, the dataset contains ground truth created from a collection of 32 films of different genres (which are not included due to copyright issues).

The violence identification is made based on two definitions of violent scenes: (1) subjective definition and (2) objective definition. The subjective definition describes a violent scene as a “scene one would not let an 8-year-old child see because they contain physical violence”.[[3]](#footnote-3) The objective definition shows that a violent scene contains “physical violence or accident resulting in human injury or pain”[[4]](#footnote-4).

# References

1. Browne, Kevin & Hamilton-Giachritsis, Catherine. (2005). *The influence of violent media on children and adolescents: A public-health approach*. Lancet. p. 8
2. National Scientific Council on the Developing Child. (2010). *Persistent Fear and Anxiety Can Affect Young Children’s Learning and Development: Working Paper No.* 9. Retrieved from [www.developingchild.harvard.edu](http://www.developingchild.harvard.edu).
3. Guerra, Nancy & Knox, Lyndee. *Classification, The Causes of Violence, Prevention and Control of Violence, Bibliography.* Retrieved 2020-05-31, from [https://law.jrank.org/](https://law.jrank.org/pages/2295/Violence.html).
4. Krug et al., *World report on violence and health*. Archived 2015-08-22 at the Wayback Machine, World Health Organization, 2002.
5. *Community Standards*, Facebook, Inc. Retrieved 2020-05-31, from [www.facebook.com/communitystandards](http://www.facebook.com/communitystandards).
6. Potter, M.C., Wyble, B., Hagmann, C.E. et al.(2014). *Detecting meaning in RSVP at 13 ms per picture*. Atten Percept Psychophys 76, 270–279. <https://doi.org/10.3758/s13414-013-0605-z>
7. Schmidhuber, Jürgen (2015). *"Deep Learning".* Scholarpedia.
8. Cireșan, D.; Meier, U.; Schmidhuber, J. (2012). *Multi-column deep neural networks for image classification*. 2012 IEEE Conference on Computer Vision and Pattern Recognition.
9. Krizhevsky, Alex; Sutskever, Ilya; Hinton, Geoffrey (2012). *ImageNet Classification with Deep Convolutional Neural Networks.* NIPS 2012: *Neural Information Processing Systems*, Lake Tahoe, Nevada.
10. "*Google's AlphaGo AI wins three-match series against the world's best Go player*". TechCrunch. 25 May 2017.
11. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei. (2009). *ImageNet: A Large-Scale Hierarchical Image Database.* IEEE Computer Vision and Pattern Recognition (CVPR).
12. Bojarski, M., Firner, B. et al. (August 17, 2016). *End-to-End Deep Learning for Self-Driving Cars*. Retrieved 2020-06-05, from <https://devblogs.nvidia.com/deep-learning-self-driving-cars/>.
13. Esteva, A., Kuprel, B., Novoa, R. et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks*. Nature* 542, 115–118. <https://doi.org/10.1038/nature21056>
14. C.H. Demarty, C. Penet, M. Soleymani, G. Gravier. (2014). *VSD, a public dataset for the detection of violent scenes in movies: design, annotation, analysis and evaluation*. In Multimedia Tools and Applications, May 2014.
15. C.H. Demarty, B. Ionescu, Y.G. Jiang, and C. Penet. (2014). *Benchmarking Violent Scenes Detection in movies*. In Proceedings of the 2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI), 2014.
16. M. Sjöberg, B. Ionescu, Y.G. Jiang, V.L. Quang, M. Schedl and C.H. Demarty. (2014). *The MediaEval 2014 Affect Task: Violent Scenes Detection*. In Working Notes Proceedings of the MediaEval 2014 Workshop, Barcelona, Spain (2014)
17. C.H. Demarty, C. Penet, G. Gravier and M. Soleymani. (2012). *A benchmarking campaign for the multimodal detection of violent scenes in movies*. *In Proceedings of the 12th international conference on Computer Vision* – Volume Part III (ECCV’12), Andrea Fusiello, Vittorio Murino, and Rita Cucchiara (Eds), Col. Part III. Springer Verlag, Berlin.
18. Olmos, R., Tabik, S., & Herrera, F. (2018). *Automatic handgun detection alarm in videos using deep learning*. Neurocomputing, 275, 66-72.

1. Foote, Keith D. (February 7, 2017). A Brief History of Deep Learning. Retrieved 2020-06-05 from https://www.dataversity.net/brief-history-deep-learning/. [↑](#footnote-ref-1)
2. <https://sci2s.ugr.es/weapons-detection>, <https://github.com/SasankYadati/Guns-Dataset>, <http://kt.agh.edu.pl/matiolanski/KnivesImagesDatabase/> [↑](#footnote-ref-2)
3. As presented in the description of the dataset. <https://www.interdigital.com/data_sets/violent-scenes-dataset> [↑](#footnote-ref-3)
4. See note 3. [↑](#footnote-ref-4)