Neurocomputing ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A mid-level video representation based on binary descriptors: A case study for pornography detection

Carlos Caetano [a,b,*], Sandra Avila [c], William Robson Schwartz [b], Silvio Jamil F. Guimarães [d], Arnaldo de A. Araújo [a]

[a] Universidade Federal de Minas Gerais, NPDI—DCC/UFMG, Minas Gerais, Brazil
[b] Universidade Federal de Minas Gerais, SSIG Group—DCC/UFMG, Minas Gerais, Brazil
[c] University of Campinas, RECOD Lab—DCA/FEEC/UNICAMP, Campinas, Brazil
[d] Pontifical Catholic University of Minas Gerais, VIPLAB—ICEI/PUC Minas, Minas Gerais, Brazil

ARTICLE INFO

Article history:
Received 20 September 2015
Received in revised form
18 March 2016
Accepted 23 March 2016

Keywords:
Binary descriptors
Mid-level representation
Bag-of-Words
BossaNova
Pornography

ABSTRACT

With the growing amount of inappropriate content on the Internet, such as pornography, arises the need to detect and filter such material. The reason for this is given by the fact that such content is often prohibited in certain environments (e.g., schools and workplaces) or for certain publics (e.g., children). In recent years, many works have been mainly focused on detecting pornographic images and videos based on visual content, particularly on the detection of skin color. Although these approaches provide good results, they generally have the disadvantage of a high false positive rate since not all images with large areas of skin exposure are necessarily pornographic images, such as people wearing swimsuits or images related to sports. Local feature based approaches with Bag-of-Words models (BoW) have been successfully applied to visual recognition tasks in the context of pornography detection. Even though existing methods provide promising results, they use local feature descriptors that require a high computational processing time yielding high-dimensional vectors. In this work, we propose an approach for pornography detection based on local binary feature extraction and BossaNova image representation, a BoW model extension that preserves more richly the visual information. Moreover, we propose two approaches for video description based on the combination of mid-level representations namely BossaNova Video Descriptor (BNVD) and BoW Video Descriptor (BoW-VD). The proposed techniques are promising, achieving an accuracy of 92.40%, thus reducing the classification error by 16% over the current state-of-the-art local features approach on the Pornography dataset.

1. Introduction

Due to the fast growth of images and videos publicly available on the Internet, the need for recognition of their contents arises. Besides the obvious need for methods related to image and video searches, it is also important to perform recognition or classification of contents that may be considered undesirable or offensive to allow the development of methods for filtering them.

The largest group of images and videos available on the Internet that people may find offensive is related to pornographic materials. A report by the ExtremeTech[1] technology site suggests that 30% of all Internet traffic is associated with pornography. They arrived at this number by estimating the traffic that a popular pornographic website generates every day and multiplied it by several other pornographic websites of similar size found on the Internet. According to their report, the largest website provider of this type of content receives three times more pageviews than major news websites (about 4.4 billion pageviews per month) and the average time spent on this site can be five times higher than in news sites.

According to Short et al. [1], pornography can be considered as any sexually explicit material with the aim of sexual arousal or fantasy. However, this definition leads to many challenges when trying to detect pornographic content, such as the bounds of "explicit" for something to be considered as pornographic material. Some works in the literature deal with this issue by dividing the material into several classes [2], complicating even more the classification task. On the other hand, there are works that choose to deal with it by using a conceptually simple evaluation considering only two classes (pornographic and non-pornographic)

* Corresponding author at: Universidade Federal de Minas Gerais, NPDI—DCC/UFMG, Minas Gerais, Brazil.
E-mail addresses: carlos.caetano@dcc.ufmg.br (C. Caetano), sandra@dca.fee.unicamp.br (S. Avila), william@dcc.ufmg.br (W.R. Schwartz), sjamil@pucminas.br (S.J.F. Guimarães), arnaldo@dcc.ufmg.br (A.d.A. Araújo).

[1] http://www.extremetech.com/computing/123929-just-how-big-are-porn-sites.

http://dx.doi.org/10.1016/j.neucom.2016.03.099
0925-2312/

[3,4], the focus of this work.

Detecting and filtering pornographic visual content from the Internet is a concern in many environments (e.g., schools and workplaces). According to Lopes et al. [5], linked text tags to pictures and videos are clearly not sufficient, since inappropriate content can be maliciously attached to seemingly innocent texts. A typical situation would be, for example, the employment of search keywords commonly used by children attached to websites with pornographic content. In addition, adults may also not wish to be exposed to such contents, for instance, from results received from search engines available on the web.

In recent years, several works in literature have been mainly focused on detecting pornographic images and videos based on visual content rather than textual information [2,3,6–15,4,16–18]. Most of these works are based on skin color detection approaches since a large fraction of pixels that have colors related to the human skin [19]. Nevertheless, a shortcoming of these approaches is related to the high rate of false positives, since not all images with large areas of skin exposure are necessarily pornographic (pictures of people wearing swimsuits, or sports-related images). Furthermore, another issue to be considered is that grayscale pictures cannot be classified using color related features.

The pornography detection task can be interpreted as a visual recognition task in the context of object recognition [5]. Approaches based on local features in conjunction with Bag-of-Words models (BoW) have been successfully applied to visual classification tasks [20,21]. In such approaches, images are represented as histograms constructed from a set of visual features. No explicit model of the object is needed and the variability of examples (related to rotation, shape scale or illumination) is treated by a training set that includes such variability. In view of that, approaches based on BoW models are suitable to the task of pornography detection.

Despite the existing methods based on BoW models produce promising results in the pornography detection context, these also make use of local feature descriptors that require a high computational processing time and generate high-dimensional real-valued vectors. For example, Avila et al. [4] made use of HueSIFT feature descriptor [22], a variant of SIFT descriptor [23] that includes color information, taking an average time of 2.5 seconds to densely extract the local features of an image generating a feature vector consisting of 165 floating point values. In fact, this is still not fast enough for real-time applications, that require a short response time. Moreover, the comparison between two extracted features would spend more computational time due to the high dimensionality. On the other hand, to satisfy the requirements of web pornographic image recognition both on precision and speed, Zhuo et al. [18] proposed a pornographic image recognition method based on the binary descriptor Oriented FAST and Rotated BRIEF (ORB), which is a low-complexity alternative. However, their work focused only on static images.

In this paper, we formalize a video descriptor approach to the visual recognition problem in the context of pornography detection in videos. The method is based on both a low-complexity alternative for feature extraction using binary descriptors and a combination of mid-level representations. We apply it to the classical BoW model generating the BoW Video Descriptor (BoW-VD). We also apply it to the BossaNova, a BoW model extension that preserves the visual information in a richer way, which generates the BossaNova Video Descriptor (BNVD). Our proposal has as advantage the fact that it does not depend on any skin detector or shape models to classify pornography; besides, according to the experimental results, it outperforms the state-of-the-art results on the Pornography dataset [4]. To the best of our knowledge, ours is the best result reported to date on the Pornography dataset employing local feature descriptors.

The use of binary descriptors and the mid-level representation for videos were first introduced in our previous works [17] and [24]. This paper presents several new aspects in comparison with the previous ones. Those aspects are highlighted in the following:

- Formalization of BossaNova Video Descriptor (BNVD). In this work, we present a new formulation which generalizes the BNVD allowing the use of different aggregation functions.
- Proposal of BoW Video Descriptor (BoW-VD). In this work, we also propose a video descriptor by using aggregation functions for combining the traditional BoW mid-level representations.
- Improvement of the experimental results. In this work, we study the behavior of binary descriptors and the mid-level representation by using several parameter settings, including the use of global pooling for creating a video descriptor. Moreover, we show the computational times for generating our video descriptor.

The remainder of this paper is organized as follows. We start by explaining the classical non-binary local feature descriptors, the most recent binary feature descriptors and the BossaNova mid-level representation (Section 2). Next, we survey the recent works on pornography detection (Section 3). We then introduce the complete formalism of our video descriptor (Section 4). Afterwards, we analyze our experiments regarding the proposed video descriptor and we perform a comparison with state-of-the-art approaches (Section 5). Finally, we present our concluding remarks (Section 6).

2. Theoretical background

The most common approach for visual recognition task consists of three distinct steps [25]: (i) extraction of local features; (ii) encoding of the local features in an intermediate representation (mid-level); and (iii) classification of the mid-level representation, usually based on machine learning techniques. Typically, the extracted local features tend to be invariant to some transformations caused by camera changes such as rotation, scale and illumination. To address these properties, local features such as SIFT [23] or SURF [26] descriptors are usually extracted. Regarding the mid-level representation, BoW models [27] are the most common approaches used to encode the extracted local features. Moreover, to improve the efficiency of retrieval and classification without sacrificing accuracy, there are several methods [28–32] that can be applied to the local features or mid-level representations in order to convert them into compact similarity-preserving binary codes. Finally, the purpose of the classification is to learn a function able to assign discrete labels to images/videos. To that end, most of the visual recognition works make use of machine learning techniques, such as Support Vector Machines (SVM).

2.1. Local feature descriptors

A local feature descriptor can be considered as a function applied to a region of the image to perform its description. The simplest way to describe a region is to represent all the pixels in this region in a single vector. However, depending on the information to be described, this would result in a high-dimensional vector leading also to a high computational complexity for a future recognition of this region [33].

In this section, we review local descriptors, which can be classified in two distinct ways [34]: (i) non-binary descriptors and (ii) binary descriptors. It is important to say that new approaches for local descriptors have been proposed in the literature, so the following list is not an exhaustive list. However, it can be

Please cite this article as: C. Caetano, et al., A mid-level video representation based on binary descriptors: A case study for pornography detection, Neurocomputing (2016), http://dx.doi.org/10.1016/j.neucom.2016.03.099