2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance

Baseline Results for Violence Detection in Still Images

Dong Wang, Zhang Zhang, Wei Wang, Liang Wang, Tieniu Tan
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
Beijing, China
{*dwang, zzhang, wangwei, wangliang, tnt* }*@nlpr.ia.ac.cn*

***Abstract*—Recognizing objectionable content draws more and more attention nowadays given the rapid proliferation of images and videos on the Internet. Although there are some investigations about violence video detection and pornographic information filtering, very few existing methods touch on the problem of violence detection in still images. However, given its potential use in violence webpage filtering, online public opinion monitoring and some other aspects, recognizing violence in still images is worth being deeply investigated. To this end, we first establish a new database containing 500 violence images and 1500 non-violence images. And we use the Bag-of-Words (BoW) model which is frequently adopted in image classification domain to discriminate violence images and non-violence images. The effectiveness of four different feature representations are tested within the BoW framework. Finally the baseline results for violence image detection on our newly built database are reported.**

***Keywords*-baseline results; violence detection; violence image database**

I. Introduction

As we enjoy the convenient and fast access to all kinds of information brought by the Internet, we are also overwhelmed by the objectionable content on it, such as violent images, bloody videos as well as pornographic information. The flourishing video-sharing websites like *Youtube* and other social websites like *Facebook* or *Twitter* even enable sharing images and videos as simple as clicking the mouse, which makes the proliferation of those harmful content possible. For those who lack proper judgment, especially young people, exposure to such negative information can lead to aggressive behavior or even crime resulting from mimicking what they see in those harmful sources [1]. In this sense, it is highly necessary to investigate the recognition of the harmful content on the Internet. In this paper, we just focus on the recognition of violence images.

Since violence detection in still images is less studied, few related works exist at present. To recognize the horror image which is one category of violecnce images, Li *et al*. propose to solve this problem by constructing the Emotional Saliency Map [2]. But this method is only designed to recognize one sub-type of violence images–*horror image*. When dealing with the violence content in a more general sense, this method may not generalize very well. Besides, there are several other methods to detect violence videos and pornographic content. Bermejo *et al*. [3] propose to use the Bag-of-Words framework and action descriptors STIP and MoSIFT to detect fighting in sports videos. Gong *et al*. [4] put forward a violence detector using low-level visual and auditory features and high-level audio effects to indentify the violent content in movies. Nam *et al*. [5] recognize the violent scenes in videos through flame and blood detection as well as motion information. What's more, Hu *et al*. [6] propose to use a C4.5 decision tree to divide the web pages into text and image pages and then recognize the pornographic content by combining text and image features. Wang *et al*. [7] use wavelet transformation and color histogram to classify pornographic images. Jiao *et al*. [8] make use of the proportion of skin area in the image and the area of the largest connected skin region as the feature vector to distinguish pornographic images.

Although there have been plenty of methods proposed in the image classification community, very few of them attempt to deal with the application of violence image detection. The first challenge facing us is how to define a violence image. A lot of ambiguity exists there because people are very subjective in distinguishing what kinds of images are indeed violent. We believe, in the context of the Internet, the concept of violence images should be extended to a more general sense under which violence images refer to not only images with people fighting, but images with gunfire, explosion, horror or bloody scenes as well. All these images can possibly arouse people's anxiety, panic and aggressiveness, which do great harm to people both physically and mentally [9]. Following this intuition, we build up a new database containing different kinds of violent images. This is one of the major contributions of our work.

Another challenge is that unlike violence detection in videos where multiple features about audio, visual and motion information can be combined to help solve the problem, detecting violence in still images can *only* rely on visual features. What's more, the violence image detection is rather complicated due to the wide range of scenes or background clutter included. In particular, variations in illumination, angle of view and dynamic backgrounds bring a lot of difficulty in coming up with an efficient detection method. Since detecting violence in still images is an image classification problem, we use the Bag-of-Words (BoW)

978-0-7695-4797-8/12 $26.00 © 2012 IEEE
DOI 10.1109/AVSS.2012.16

model which is widely adopted in this domain as our framework. In terms of feature representation within the BoW model, we test four different features to evaluate their effectiveness in discriminating the violence images from non-violence images.

The remainder of this paper is organized as follows: In Section 2, we introduce the new established violence image database. Section 3 presents our approach for violence detection in still images and summarizes the empirical results of our approach. Section 4 concludes this paper.

II. Database

A proper database is essential to evaluate the approaches for solving the task of violence image detection. To date, unfortunately, we have not found a widely used or publically-available database about violence image detection. So we decide to establish a new database on our own, named as VID (Violence Image Detection) database here.

Initially we collected around 1000 violence images and 2000 non-violence images in the database. Most of the images mainly come from the online searching engines like *Google* ,*Yahoo* and *Baidu*. Query words such as "violence ", "horror ", fight", "explosion" ,"blood", "gunfire" and so on were used when collecting these images. Some others are the screenshots of violent movies or video clips. They are all color images in JPEG format with height no larger than 1000 pixels and width no larger than 500 pixels. Given our definition of violence images in the introduction part, this database includes images about fierce fighting or conflicts in sports events, explosion, gunfire and bloody or horror scenes, etc. Although such a database may be difficult to classify due to the relatively large within-class difference, it can capture a full variety of violence image categories.

As discussed before, when collecting violence images people behave very subjective so that there is inevitable bias in this database. To reduce this kind of bias or subjectivity, we ask 20 PhD volunteers in our lab to manually label the 3000 candidate images according to the degree of violence (such as fierceness of fighting, how horror or brutal an image is). In detail, each candidate image can be assigned three different labels, i.e, "Violence", "Non-violence" and "Neutral". When the annotator has high confidence to confirm a candidate image as "Violence" or "Non-violence", the candidate images are assigned the corresponding labels directly. If the annotator can not clearly tell whether this image contains violent information or not, then we label it as "Neutral". Finally we choose the top 500 images with the most "Violence" labels as final violence image samples and the top 1500 images with the most "Non-violence" labels as final non-violence image samples, whose examples are shown in Fig. 1 and Fig. 2 respectively. This resulting database enables us to evaluate the performance of our approach for violence detection in images, which also provides assistance and convenience for further research in this area. The database will be available by request.

Figure 1. Samples of violence images in the VID database.

Figure 2. Samples of non-violence images in the VID database.

III. Baseline Algorithm and Experimental Results

In this section, we introduce the framework of our approach for classifying violence and non-violence images. Although there are some classification methods specificly targeting some type of violence image such as the horror image, no other methods exist for discriminating the violence images with many different varieties from the non-violence ones. So here we provide our baseline algorithm for further research and comparison in this area.

A. The Bag-of-Words (BoW) Model

The Bag-of-Words (BoW) model, which is used to be a order-free document representation in Natural Language processing (NLP), has been widely adopted as a main framework for computer vision tasks such as image classification. The BoW model represents each image through a histogram over a bunch of *visual words* in a visual dictionary (codebook), which corresponds to the number of occurrences of particular image patterns in a given image [10]. While constructing the codebook, the visual words in it are usually defined as the cluster centers generated from the K-means clustering over a pool of low-level feature descriptors such as SIFT [11]. The BoW model is favored

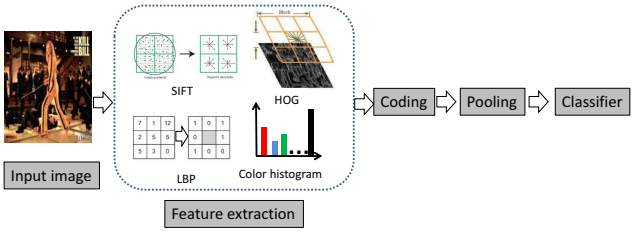

Figure 3. The pipeline of Bag-of-Words model

by the image classification community due to its simplicity, computational efficiency and robustness to occlusion and within-class variance. As illustrated in Fig. 3, the BoW model usually consists of three major procedures, which are feature extraction, feature coding and feature pooling.

Since the BoW model represents an image as an orderless collection of local features, it discards the information about the spacial layout of the features and thus has limited descriptive ability. Specificly, it can not describe shape or segment an object from its background. To overcome the disadvantage of the basic BoW model, Lazebnik *et al.* [12] proposed a method called *Spatial Pyramid Matching* (SPM) that repeatedly subdivides the image and computes histograms of local features at increasingly fine resolutions. Therefore, the spatial information can be encoded in the BoW model. Besides there are usually two kinds of coding methods in the BoW model, namely hard voting and soft voting. The hard voting strategy represents the feature with its nearest visual code while in soft voting the feature is represented as the weighted average of all visual words rather than only one code. Hard voting is inferior to soft voting since it does not consider the codeword ambiguity and often produces quantization error. So we prefer soft voting to be our coding scheme in the BoW model.

Among the four basic procedures in the BoW model, we pay the most attention to feature representation so as to evaluate the effectiveness of different features in classifying violence ans non-violence images. Four commonly used features are chosen as a comparison: SIFT, HOG [13], LBP [14] and color histogram. These four features are briefly introduced as follows:

SIFT: The SIFT features are invariant to image scale and rotation, and can provide efficient matching across a certain range of affine deformation, view-angle change and illumination difference. It's one of the most commonly adopted features in object recognition.

HOG: The HOG descriptors share some similarities with the SIFT descriptors, but they are computed on dense grid of uniformly spaced cells and they use local contrast normalizations to improve their robustness to illumination change.

LBP: A s a feature for texture analysis, LBP labels the pixel of an image by thresholding the neighborhood of each pixel with the value of the center pixel and considers the result as a binary number. Because of its discriminative power and computational simplicity, LBP texture descriptor has become a popular approach in various applications such as face recognition.

Color histogram: A color histogram is a statistic that represents the distribution of colors in an image. It is relatively invariant with translation and rotation about the viewing axis. In particular, the color histogram is suitable for recognizing an object of unknown position and rotation in a scene. For our problem of violence image recognition, color may be a useful feature for classification given there are some bloody or explosive scenes involved in our violence image database.

To sum up, in order to recognize violence and non-violence images, we adopt the BoW model integrated with the SPM scheme and soft voting strategy. And four different feature representations are tested in the following experiments.

B. Experimental Setup

In the experiments, the codebook in the BoW model is constructed via K-means clustering and the number of codes is set to be 8192 in the following experiments. SPM is performed on three levels which are 1×1, 2×2, 3×1. For other feature representations, we follow the default settings in the original literatures. During the classification stage, the Lib-linear SVM [15] is chosen to be the classifier. For our new established dataset, we randomly select one half of the images as the training set while keep the rest images as the testing set. The experiments are repeated 10 times and the average classification accuracy and standard deviation are finally reported.

C. Experimental Results and Analysis

The baseline results for our experiments are demonstrated in Table 1 and we give our analysis afterwards.

Table I
VIOLENCE IMAGE RECOGNITION USING DIFFERENT FEATURES

Adopted Feature	Average classification accuracy
1. BoW+SIFT	$85.7 \pm 1.4\%$
2. BoW+HOG	$84.3 \pm 1.6\%$
3. BoW+LBP	$90.1 \pm 1.5\%$
4. BoW+Color histogram	$84.1 \pm 1.3\%$

As we can see, among the four different feature representations, LBP has achieved the best classification performance, followed by SIFT and HOG. On the contrary, color histogram turns out to be the least discriminative in classifying violence and non-violence images.

The above results manifest that texture descriptors like LBP could be more effective and suitable in our task. The new dataset built on our own is rich in texture and somehow difficult to classify due to the large within-class difference

and complex background clutters. HOG and SIFT perform poorly when the background is cluttered with noisy edges. But LBP is complimentary in this aspect since it can filter out noises using the concept of uniform pattern.

For color histogram, it does not perform as promising as what we expect it to be. The reason is that the representation is only dependent on the color distribution of the image, ignoring the shape and texture information of the objects in the image. Color histogram can possibly be identical for two images with different object content which happens to share the same color information. In other words, two images with different semantics could be considered similar if they have similar color distribution.

IV. Conclusions

Violence detection in still images is becoming increasingly important since it can help prevent the proliferation of objectionable content on the Internet. Motivated by this problem, we have established a new violence image database covering a large variety of violent sources, which will facilitate further research on violence image detection. Besides, we adopt the BoW model to solve the violence image classification problem. Four different feature representations are tested in the experiments and the texture descriptor LBP has demonstrated the most discriminative capability. Thus our method provides a baseline result for violence detection in still images.

In terms of future work, we will continue to refine and enlarge our violence image database by collecting more diversified violence images. One possible way is through searching different well-acknowledged violence image queries via online image searching engines. In this way we can obtain a large amount of violence images while saving much human labor. Another future work we may care about is to exploit some other effective feature representations and fuse multiple-cue features so that a more accurate and robust violence image detection algorithm can be developed.

Acknowledgment

This work is funded in part by the the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDA06030300), the National Basic Research Program of China (Grant No.2012CB316300), Hundred Talents Program of Chinese Academy of Sciences and National Science and Technology Support Program (Grant No.2011BAH11B01).

References

[1] C.-L. P. L. Rowell Huesmann, Jessica Moise-Titus and L. D. Eron, "Longitudinal relations between childrens exposure to tv violence and their aggressive and violent behavior in young adulthood: 1977–1992," *Developmental Psychology*, vol. 39, pp. 201–221, 2003.

[2] B. Li, W. Hu, W. Xiong, O. Wu, and W. Li, "Horror image recognition based on emotional attention," in *Proc. of 10th Asian Conference on Computer Vision (ACCV)*, 2010, pp. 594–605.

[3] E. B. Nievas, O. D. Suarez, G. B. García, and R. Sukthankar, "Violence detection in video using computer vision techniques," in *Proc. of the 14th International Conference on Computer Analysis of Images and Patterns - Volume Part II*, 2011, pp. 332–339.

[4] Y. Gong, W. Wang, S. Jiang, Q. Huang, and W. Gao, "Detecting violent scenes in movies by auditory and visual cues," in *Proc. of the 9th Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing*, 2008, pp. 317–326.

[5] J. Nam, M. Alghoniemy, and A. H. Tewfik, "Audio-visual content-based violent scene characterization," in *Proc. of IEEE International Conference on Image Processing*, 1998, pp. 353–357.

[6] W. Hu, O. Wu, Z. Chen, Z. Fu, and S. Maybank, "Recognition of pornographic web pages by classifying texts and images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 1019–1034, 2007.

[7] J. Z. Wang, J. Li, G. Wiederhold, and O. Firschein, "System for screening objectionable images," *Computer Communications Journal*, vol. 21, pp. 1355–1360, 1998.

[8] F. Jiao, W. Gao, L. Duan, and G. Cui, "Detecting adult image using multiple features," in *Proc. of IEEE International Conference on Image Processing*, 2001, pp. 378–383.

[9] B. J. Bushman and L. R. Huesmann, "Short-term and long-term effects of violent media on aggression in children and adults," *Archives of Pediatrics and Adolescent Medicine*, vol. 160, pp. 348–352, 2006.

[10] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *In Workshop on Statistical Learning in Computer Vision, ECCV*, 2004, pp. 1–22.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–100, 2004.

[12] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *In CVPR*, 2006, pp. 2169–2178.

[13] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *In CVPR*, 2005, pp. 886–893.

[14] T.Ojala, M. Pietikinen, and D.Harwood, "A comparative study of texture measures with classification based on feature distributions," *Pattern Recognition*, vol. 29, pp. 51–59, 1998.

[15] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.