International Journal of Computer Applications Technology and Research
Volume 12–Issue 03, 14-17, 2023, ISSN:-2319–8656
DOI:10.7753/IJCATR1203.1004

A Two-stream Convolutional Neural Network-based Pornography Recognition Method

Congcong He
School of Communication Engineering
Chengdu University of Information Technology
Chengdu, China

Qinglin Huang
School of Communication Engineering
Chengdu University of Information Technology
Chengdu, China

Jieyuan Luo
School of Communication Engineering
Chengdu University of Information Technology
Chengdu, China

Abstract: The main approach taken to identify pornographic video content is achieved by performing pornography detection on the video content. By extracting features from video key frames and using some common neural network models to recognize the extracted key frame images, a certain accuracy rate can be obtained. However, another key information of video recognition, action information, is ignored, which leads to misclassification of some indistinguishable videos such as sumo wrestling and boxing. A dual-stream convolutional neural network-based pornographic video recognition method is proposed to address this problem. The experimental results show that the dual-stream convolutional neural network effectively improves the recognition rate of indistinguishable pornographic videos.

Keywords: Video Identification, Two-stream convolutional neural network, Keyframe styling

1. INTRODUCTION

With the rapid development of the short video and live streaming industry, the audience of short video and live streaming is becoming more and more widespread. Many primary and secondary school students like to watch live or short video, video content safety issues are very serious. At present, many Internet companies still use human supervision for video supervision. In this period of short video screens everywhere, human supervision consumes a lot of human, material and financial resources. At the same time long time human supervision is also a great threat to the psychological health of the regulator This paper proposes the use of dual-stream convolutional neural network model to improve the recognition efficiency and reduce the misjudgment rate of difficult-to-identify videos.

2. DEVELOPMENT OF NEURAL NETWORK

In 1996, D. A. Fprsyth and M. Fleck successfully implemented a nude recognition system by studying the color and texture characteristics of skin tones. m.-H. Yang and N. Ahuja et al. used the distribution of skin pixels in color space for modeling and used the model for detecting skin tone acorns. m. J. Jones and J. M. Rehg et al. The histogram of color distribution was derived from RGB color space, and then the histogram of color distribution of normal images was compared with the histogram of color of pornographic images, and convolutional neural network was used to classify normal images from pornographic images, and finally, pornography identification was achieved.[1]

The above methods, the effect of pornographic image identification based on the detection of skin color depends only on skin color pixels. These methods are too single in judging the labeling and not highly reliable.In 2012, the AlexNet convolutional neural network model was introduced and its designer Alex Krizhevsky won the ImageNet large-scale vision challenge using the AlexNet convolutional neural network model. the shockingly high recognition rate of AlexNet has led many scholars to join the research of convolutional neural networks. For video recognition, Karen Simonyan et al. used dual-stream convolutional neural networks for two dimensions of temporal and spatial information to study the characteristics of video that are different from images. Based on many studies on image recognition and video recognition, this paper proposes a method to recognize pornographic videos using Two-Stream CNN model. Since the key frame image only contains the spatial information of the image, the action information of the video is completely lost, which cannot achieve the role of video recognition, and the final result will not be able to recognize some indistinguishable pornographic videos well, causing a large rate of misjudgment. Adding video action information to the model, optical stream can express the change information of adjacent images, and the pornography recognition method based on Two-Stream CNN has good reliability[2].

3. CONVOLUTIONAL NEURAL NETWORK

3.1 Convolutional neural network model

Yann LeCun, a tenured professor at New York University, proposed the convolutional neural network in order to recognize handwritten letters. In the convolutional neural network has developed rapidly for several years, but its structural principle is roughly the same as that originally proposed by Yann leCun, consisting of a convolutional layer, a downsampling layer, and a classification layer[3]. As shown in Figure 1.

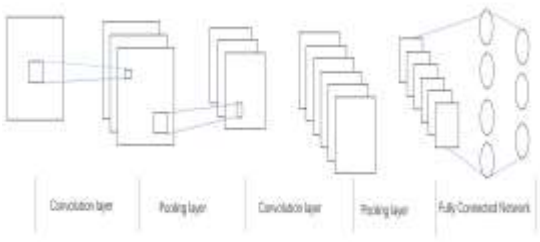

Figure 1. CNN typical struture.

CNN is to learn the features of the image by convolutional operation, the convolutional layer receives the output data of the data downsampling layer and multiple convolutional kernels for convolution, the formula of convolutional operation is.

$$Y_{i,j}^{l} = f(\sum\nolimits_{i\in m_h}\sum\nolimits_{i\in n_w} x_{i,j}^{l} G_{i,j}^{l} + b^{l})$$

In the above equation, $Y_{i,j}^{l}$, denotes the value of the (i,j) point output from the Lth layer, $x_{i,j}^{l}$ denotes the input value of the (i,j) point in the Ist layer, G denotes the convolution kernel, b denotes the bias term, m_h, n_w denotes the window size of the local perceptual field in the Lth layer.[4]

In the structure of CNN, multiple feature maps are output after one convolution. Downsampling the feature maps makes the network robust to image rotation, translation, and scale transformation, and reduces the computational effort of network training.[5] Commonly used downsampling methods are mean sampling and maximum sampling. The mean sampling averages the feature values in the sampling window as the sampling result, and the mean sampling formula is:

$$Y_{i,j} = \max_{0\leq h\leq H-1, 0\leq w\leq W-1} (x_{i*H+h, j*W+w})$$

In the above equation, $Y_{i,j}$ is used as the output value of the convolutional neural network downsampling, X_(1*H+h,j*w+w) is used as the input value of the convolutional neural network, and H,W denotes the length and width of the sampling window. The maximum downsampling method is to take the maximum value within the sampling window as the sampling result, and the calculation formula is.

$$Y_{i,j} = \max_{0\leq h\leq H-1, 0\leq w\leq W-1} (x_{i*H+h, j*W+w})$$

3.2 Two-stream network model

Each video contains both temporal and spatial information. The change of background from each image frame to the next represents the temporal information of the video. The background of the behavior in the video represents the spatial information of the video. For extracting the number of key frames, passing each of these frame images through a neural network does this very poorly. the Two-stream network design compensates for these deficiencies. Both spatial and temporal streams are placed on the Two-stream network model. The video extracted keyframe images and optical flow images are trained on the Two-stream network model, and the extracted spatial and temporal information is fused using the mean fusion method for the recognition results.

The extracted keyframe images are fed into the Two-stream network as spatial information. The keyframe images are sufficient for the relatively easy to distinguish pornographic videos. The extracted optical flow images are fed into the Two-stream network as motion information[6].

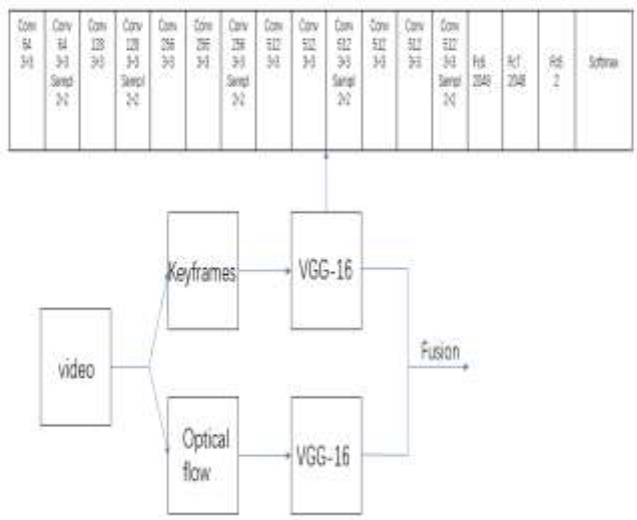

Figure 2. Two-Stream Model Architecture.

3.3 Video Segmentation

Firstly, the video should be segmented, and then the segmented video should be extracted with key frames and optical flow images. In this paper, we mainly refer to the method of video segmentation with unsupervised clustering proposed by Jin Hong et al. Firstly, the video is divided into n classes using clustering method, and the video frames in each class have similarity, and the frames in different classes are not similar or have low similarity. For the class with fewer frames, i.e., the class is not representative, it is directly merged with the neighboring frames[7].

The images stored in the computer are in RGB space, and since HSV color space has better color characteristics than RGB color space, it is necessary to map the colors to HSV space[8]. The preprocessing steps are as follows:

1. Map the RGB images distributed in 0~255 directly to the HSV color space of 0~255.

2. partition the HSV colors, divide the H component into 12 equal blocks, and the S and V components into 5 equal blocks each. 3. map the original colors in the range of 0~255 to the range of 12×5×5.

4. Build the color space of HSV, set the image size as M×N, and count the percentages of H, S, and V components respectively.

The calculation formula is shown in 3.1:

$$H(i) = \frac{H\_follow(i)}{M * N}$$

$$S(j) = \frac{S\_follow(j)}{M * N}$$

$$V(k) = \frac{V\_follow(k)}{M * N}$$

To calculate the similarity of two images, we need to calculate the similarity of the three color histograms H,S,V, respectively, by the minimum value corresponding to the same index of the histogram of the two images is accumulated. The formula is as follows:

$$S_h(f, Shot) = \sum_{i=1}^{12} \min\left(H(i), Shot\_H(i)\right)$$

$$S_s(f, Shot) = \sum_{i=1}^{5} \min\left(S(i), Shot\_S(j)\right)$$

$$S_v(f, Shot) = \sum_{i=1}^{5} \min\left(V(k), Shot\_V(k)\right)$$

Since the human eye is more sensitive to the H component than to the S component, and the S component is greater than the V component, for each.

The weights are set for each component, 0.5 for H, 0.3 for S, and 0.2 for V.

The video segmentation and key extraction algorithm using clustering is described as follows:

Maintain a center of mass for each class.

1.For each frame, use Equation 3.2 to calculate the similarity of the cluster cores, if the similarity is less than the set valve value, then it will be placed in a new class, otherwise it will be added to the previous class.

2.Merge some of the clusters that are too small.

3.According to the clustering results Class1,... , Class, perform video segmentation.

4.Calculate the first and largest image in each cluster and use it as a key frame.

3.4 Extraction of optical flow diagrams

Optical flow is a technique that is widely used in computer vision and computer graphics. The concept was first introduced by Gibson in 1950. Optical flow describes pixel motion information, such as the direction of motion as well as the speed of motion. The change of pixels in the image sequence in the time domain and the correlation between adjacent frames are used to find the correspondence that exists between the previous frame and the current frame. In general, the motion information in a video is mainly generated by the motion of the foreground target, the motion of the camera, or the joint motion of both, and is described by optical flow. Optical flow algorithms can be applied to many fields, such as video processing, robot vision, virtual reality, etc. In visual perception, when the human eye observes a moving object, the moving object leaves a continuous image on the retina, just like the flow of light; therefore, the motion information in the video is called optical flow[7].

Optical flow is further divided into sparse optical flow and dense optical flow. Sparse optical flow is a type of image alignment method that specifically targets sparse points on an image, that is, given a number of points on a reference map, these points are generally corner points, and find their counterparts in the current image. Dense optical flow is an image alignment method that matches point by point for an image. In this paper, we mainly use dense optical flow to extract action information from video. The dense optical flow obtains the motion information in the video by performing a complete calculation of the offsets of all points on the image. It can be seen that the computation of dense optical flow is much larger than that of sparse optical flow, so the effect of its alignment is significantly better than that of sparse optical flow alignment[8].

his paper uses the Lucas-K anade optical flow algorithm, which is a common optical flow algorithm in Opencv. It has three assumptions: firstly, the luminance is constant. Secondly, the image motion varies little with time. Finally, spatially consistent[9].

The constraint equation of the image is :

$$I(x, y, z, t) = I(x + \Delta x, y + \Delta y, z + \Delta z, t + \Delta t)$$

4. ANALYSIS OF EXPERIMENTAL RESULTS

4.1 Data pre-processing

This paper uses the NPDI dataset, which is a public pornography dataset containing a total of 800 videos, including 400 pornographic videos and 400 non-pornographic videos, with a total duration of 77 hours. The videos are divided into 200 videos that are easily distinguishable and 200 videos that are not easily distinguishable. The easily distinguishable videos are mainly about eating and playing games. The content of the videos that are not easily distinguishable is mainly swimming, wrestling, etc.

All the videos in the NPDI dataset were keyframed, and on average, multiple keyframes were extracted from each video, and a total of 16727 keyframes were extracted from all 77 hours of video frames. The entire dataset is described in Table.[10]

Table 1. Dataset

Category	Number of Videos	Duration/h	Frames Per Video
Pron	400	57	15.6
Non-Porn(easy)	200	11.5	33.8
Non-Porn(difficulty)	200	8.5	17.5
All Videos	800	77	20.6

4.2 Experiment content

NPDI consists of 16727 keyframe images extracted from 800 videos, of which 10340 are non-pornographic images and 6387 are pornographic images. The non-pornographic images are further divided into 6785 easy-to-distinguish images and 3555 hard-to-distinguish images. In the training, 80% of the pre-processed images are used as the training set, and the top, bottom, left, right and middle images are extracted and then flipped horizontally by mirroring, and 10 training images are generated for each image.

The training of optical flow convolutional network extracts dense optical flow images from 16727 segmented videos in the NPDI dataset. In this paper, the stacking unit is 10, and the dimension of the stacked optical flow image is (224,224,20), where 20 can be regarded as the number of channels of the image. A total of 151,893 optical flow stacking data are obtained after stacking all extracted optical flow images, and the label of each optical flow stacking data is consistent with the classification label of the corresponding video[11].

In order to evaluate the detection effect of Two-stream model, 50 pornographic videos and 50 non-pornographic videos were randomly selected from the NPDI dataset to test the VGG16 model and Two-stream model, and the accuracy, recall and F1 were calculated for each model. The higher the classification effectiveness of the models. The test results are shown in the table below. Comparing the data in the table, we can conclude that the VGG-16 model has a higher accuracy and recall rate than the Two-stream model, and is more accurate in detecting pornographic videos.

Table 2. Experiment results

Model	Accuracy(%)	Reall(%)	F1(%)
VGG-16	93.3	93.1	93.2
Two-Strean	95.2	95.0	95.1

To address the problem that traditional recognition of keyframe images leads to a high false positive rate for indistinguishable videos such as swimming, the Two-Stream model was tested using the same test data set for both models. The test data included 10 boxing videos, 10 swimming videos, and 10 breastfeeding videos. The test results indicated that the M-Two-Stream model improved the classification of videos such as boxing from 50% to 80% compared to the VGG-16 model, the 30% accuracy for the swimming category to 80%, and the 10% accuracy for the lactation videos, showing through experiments that the Two-Stream model reduced the misclassification rate for difficult video detection.

5. CONCLUSION

In this paper, a two-stream convolutional neural network based pornographic recognition method (Two-Stream CNN) is proposed. The method introduces motion information in the video, video segmentation, key frame extraction, feature extraction, and feature combination. The experiments show that the Two-Stream model has higher accuracy in detecting pornographic videos compared with the VGG-16 model, and reduces the misclassification rate of hard-to-distinguish videos such as breastfeeding, wrestling, swimming, and sumo wrestling.

6. REFERENCES

[1] D.A.Forsyth, 1996.M.Fleck.Finding naked people[C]. In Proc.European Conference on Computer Vision.

[2] M-H. Yang，N.Ahujia.1999. Gaussian mixture model for human skin color and its application in image and video. database[J]. SPIE Storage and Retrieval for Image and Video Database.

[3] M.J.Jones, JM. Rehg.2002.Statistical color models with application to skin detection[J]. International Journal of Computer Vision.

[4] Srisaan C.A classification of internet pornographic images[J]. International of Electronic Commerce Studies.

[5] Basilio J AM, Torres G A,Perez G S, et al.2011. Explicit content image detection[J]. Signal & Image Processing,

[6] Karavarsamis S, Atarmos N,Blekas K, et al.2013 Detecting pornographic images by localizing skin ROIs[J].International Journal of Digital Crime & Forensics.

[7] Krizhevsky A, Sutskever I, Hinton G E.2012. Imagenet classsificaton with deep convolution neural networks[C].Advances in Neural Information Processing Systems. Lake Tahoe:NIPS.

[8] Simonyan K,Zisserman A.2014. Very deep convolutional networks for large-scale image recognition [J].Computer Science.

[9] Szegedy C, Liu W, Jia Y, et al.2015. Going deeper with convolutions[C]. Proc of TEEE Conf on Computer Visionand Pattern Recognition. IEEE.

[10] He kaiming, Zhang Xiangyu, Ren Shaoqing, et al.2015 Deep residual learning for image recognition[J]. ComputerScience.

[11] Mohamed N. 2015.Moustafa.Applying deep learning to classify pornographic images and videos[C]. Pacific-rimSymp on Image and Video Technology.