Deep feature learnt by conventional deep neural network

Author:
Niu, H; Xu, W; Akbarzadeh, H; Parvin, H; Beheshti, A; Alinejad-Rokny, H

Publication details:
Computers and Electrical Engineering
v. 84
0045-7906 (ISSN); 1879-0755 (ISSN)

Publication Date:
2020-06-01

Publisher DOI:
https://doi.org/10.1016/j.compeleceng.2020.106656

License:
https://creativecommons.org/licenses/by-nc-nd/4.0/
Link to license to see what you are allowed to do with this resource.

Downloaded from http://hdl.handle.net/1959.4/unsworks_69046 in https://unsworks.unsw.edu.au on 2024-04-27

Computers and Electrical Engineering 84 (2020) 106656

Contents lists available at ScienceDirect

Computers and Electrical Engineering

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

Deep feature learnt by conventional deep neural network[☆]

Huan Niu[a], Wei Xu[b,c,*], Hamidreza Akbarzadeh[d], Hamid Parvin[d,e,*], Amin Beheshti[f], Hamid Alinejad-Rokny[g,h,i,*]

[a] *School of Information and Communication Engineering, Communication University of China, Beijing 100000, PR China*
[b] *Department of Information Technology, Hubei University of Police, PR China*
[c] *Hubei Collaborative Innovation Center of Digital Forensics and Trusted Application, PR China*
[d] *Departeman of Computer Science, Nourabad Mamasani Branch, Islamic Azad University, Fars, Mamasani, Iran*
[e] *Young Researchers and Elite Club, Nourabad Mamasani Branch, Islamic Azad University, Fars, Mamasani, Iran*
[f] *Department of Computing, Macquarie University, Sydney, 2109, Australia*
[g] *Systems Biology and Health Data Analytics Lab, the Graduate School of Biomedical Engineering, UNSW Sydney, Sydney, 2052, Australia*
[h] *School of Computer Science and Engineering, the University of New South Wales (UNSW Sydney), Sydney, 2052, Australia*
[i] *Health Data Analytics Program Leader, AI-enabled Processes (AIP) Research Centre, Macquarie University, Sydney, 2109, Australia*

ARTICLE INFO

Article history:
Received 8 May 2019
Revised 15 March 2020
Accepted 18 March 2020
Available online 20 April 2020

Keywords:
Intelligent filtering system
Image classification
Deep feature
Deep neural network
Convolutional neural network
Data analytics

ABSTRACT

In this paper, we introduce an approach to discriminate unconventional images and their intelligent filtering. As the target data to this issue are huge and consequently, a handling approach might potentially be a very time consuming one, one of the major challenges to be solved by this introduced approach is its ability for dealing with large-scale datasets. A deep neural network might be a good option to resolve this challenge. It can provide a good accuracy while dealing with huge databases. In the proposed approach, the new architecture is introduced using a combination of AlexNet and LeNet architectures. It uses convolutional, polling and fully-connected layers. The results are tested on two large-scale datasets. These tests show that the introduced architecture is more accurate than the other recently developed methods in identifying unconventional images. The proposed approach may be used in different applications such as intelligent filtering of unconventional images or medical images analysis.

© 2020 Elsevier Ltd. All rights reserved.

1. Introduction

Sharing of informative resources such as images and their worldwide distribution are increasing daily by fast growth of internet and web social media. Although this phenomenon has many advantages such as development of content sharing in educational and news areas, it has raised some concerns in this area too. There are concerns like distribution of unwanted, offensive and porn images on internet websites. This problem shows the importance of unconventional images filtering. Also, there are many social sites that they have a small percentage of unconventional images. In the absence of suitable intelligent filtering, the government is forced to filters the whole sites. According to the vast usage of internet users, the feeling of dissatisfaction is created among internet users. In fact, employing an intelligent filtering enables internet users to access many sites. But it limits users to access some sites such as porn images. In this paper, we present an approach for identification of unconventional images that is important in the intelligent areas.

[☆] This paper is for regular issues of CAEE. Reviews processed and recommended for publication to the Editor-in-Chief by Area Editor Dr. E. Cabal-Yepez.
[*] Corresponding authors.
E-mail addresses: wei2016@outlook.com (W. Xu), parvin@iust.ac.ir (H. Parvin), h.alinejad@ieee.org (H. Alinejad-Rokny).

https://doi.org/10.1016/j.compeleceng.2020.106656
0045-7906/© 2020 Elsevier Ltd. All rights reserved.

Using internet as an important reference for students in the education usually has many risks for society. There are risks such as access of kids to porn images. According to above mentioned cases, legal constrains are usually applied to access to unconventional images for individual under 18 years (legal age) [1]. With regard to the widespread usage of this information, applying constrains cannot solve all the related problems. Therefore, in recent years, many research studies have employed image processing for designing a filtering system in order to intelligent filtering of unconventional images. We are unable to identify and separate unconventional images easily, due to the nature of image structure and their complexity.

In some of related studies, skin color and related features have been used for identification of unconventional images. The first step in recognition of these images is detection of pixels associated with skin color. But these pixels are not trusted enough for identification of unconventional images [2].

Considering the above mentioned issues, it is very important to design a powerful system for analyzing images and their intelligent filtering. On the other hand, in recent years, deep neural networks have been used to identify and recognize images in a large data set [3]. Therefore, in this paper, a comprehensive system based on deep neural network learning with a new structure has been introduced for intelligent filtering of unconventional images. This new structure is fully described in Section 3.2. The methods presented in this study have been evaluated on extracted color images of different web pages. Also, the results of new methods are reported on color images. But most of these results can also be evaluated on gray images. The results show a better performance of this method than the other state of the art methods.

In the second part of this article, we will introduce the related work in the field. In the third section, deep neural networks and the proposed network architecture are expressed. In the following, we introduce the data set, performed tests and also their analysis. Finally, the conclusion of the paper and future work are presented.

2. Related work

In recent years, many studies have focused on identification of unconventional images. The filtering of unconventional images based on content and intelligent filtering are of the most important topics in these studies. In general, statistical classifiers [4] and statistical image detectors [5] are used in this type of filtering. In these classifiers, images are divided into two categories: conventional and unconventional images. In general, previous studies for recognition of unconventional images are divided into four categories: (a) methods based on detection of human body structure, (b) methods based on image retrieval, (c) methods based on visual words and (d) methods based on skin features. In the following, we introduce and describe the related work based on this categorization.

Methods based on the human body structure have been emerged in references like [6] since 1996. In these methods, areas associated with skin color are detected at first. Then, depending on body structure information and extractive areas of skin color, areas related to human body are recognized. Finally, the unconventional images are recognized according to areas extracted from human body. It is difficult to find a comprehensive model of human body anatomy, considering the non-rigid nature of human body and various states of body in images. A good model for human body structure has not yet been presented. Therefore, methods based on human body structure have low accuracy.

The purpose of image retrieval based methods is to obtain the best visual matching for identification of subclasses of dataset. In this method, the unconventional images matching input images are extracted from database. If the number of matching unconventional mages exceeds the specific threshold, the input image is recognized as unconventional image. In these methods, image recognition is largely dependent on database. According to the frequency of conventional and unconventional images in different shapes, it is difficult to create a complete database. In these methods, in order to improve the recognition accuracy, it is essential to use a large number of samples which reduces the identification time at the test stage [7].

In recent years, the researchers have focused on the identification of images based on visual words. This method has attained many attentions and has become popular in classification of objects recently [8]. According to the definition [8], visual words refer to the small pieces of images which contain information related to color, object or texture. The researchers have presented a method for image analysis inspired by text content analysis. According to these methods, an image is considered as a combination of words; therefore, the content analysis used in semantic annotation is used for recognition of unconventional images. In these methods, the visual words are extracted in order to describe semantic content of images; then bag of word (BoW) model is presented for recognition of unconventional images. The visual words and semantic information of associated tags are used in order to estimate the relationship level between BoW model for recognition of tags available in image and unconventional images.

In the methods based on skin features, the recognition of images can be considered as a classification task [9]. According to these methods, color and texture features are extracted from the skin; then pattern classifiers are used in order to recognize and classify images [9]. Many of recent related works have been dedicated to these methods. These methods are divided into three general categories: (a) methods based on color information, (b) methods based on shape information, and (c) methods based on descriptors of local features. Methods based on (pixel) color features are used in recognition of skin areas. In reference [10], more than 200,000 images are selected from conventional and unconventional sites; the amount of hue in conventional images is very different to the amount of hue in unconventional images (Fig. 1). Every pixel of image can be classified as a skin or non-skin pixel. The methods usually classify pixels by SVM classifier [9], multi-layer perceptron (MLP) artificial neural networks (ANN) classifier [10] and decision tree classifier [7].

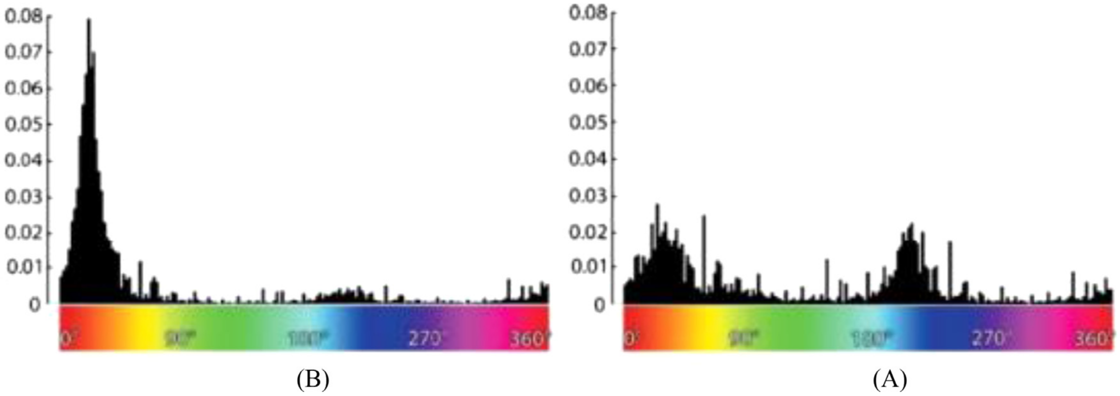


Fig. 1. The comparison between the results obtained from the distribution of the amount of hue in images extracted from conventional and unconventional websites [10], the normalized value of the hue by an angel between 0 and 360° is calculated according to the definition of the HSV color space. (A) Conventional images, (B) unconventional images.

The second category of intelligent filtering methods based on skin is the category of methods based on shape information [10]; these methods use handmade features [4] and not automatically extracted features [11]. These methods focus on recognition of skin areas. Also, a number of features based on image area descriptors have been used in these methods. The features used in this category can be divided into five subcategories including (b.1) features based on contour, (b.2) features based on moments, (b.3) features based on geometric constrains, (b.4) features based on color segments and (b.5) features based on MPEG features.

The third category of intelligent filtering methods based on skin contains the methods which use the feature vector to describe an area of image. In these methods, key points are selected by sparse sampling of images. These key points are used as features for recognition of similar images. SIFT [12] and probabilistic latent semantic analysis (PLSA) [13] can be considered as two researches associated with these methods.

The main challenge in all these methods is to extract an effective and useful feature set for discrimination of conventional and unconventional images [8]. At first, the features of images are extracted then classification of images is done. The classifier performance is depended on features extracted in its previous stage. In this paper, the machine learning with deep convolutional neural networks has been used to extract the features and classification simultaneously. In this paper, intelligent identifications and filtering of unconventional images based on DNN has been done. Therefore, the next section is dedicated to DNN.

3. Deep neural network

Nowadays, the ANNs are used as classifier in any classification task since they have suitable capability for learning new and combined features. The deep neural network (DNN) is a model of ANN family that is used for learning of nonlinear transformation on data. The most important difference between a DNN and a regular ANN is its capability in recovering data in each layer. In another word, the DNN keeps features space information by modeling the distribution of data in each layer. The DNNs have more generalization power on test data and there is no overfitting in the DNNs. In fact, in the regular ANNs, layers with full connectivity are wasteful. On the other hand, the large number of parameters in the regular ANNs increases the possibility of overfitting. However due to use of weight sharing methods, local receptor field, spatial down sampling, the probability of overfitting is reduced in the DNNs.

The extraction of meaningful hand-crafted features is difficult in some applications. Therefore, the extraction of features based on learning is so important. DNNs have become very popular due to their flexible design and formulas to solve different problems and also they have been applied particularly in the medical image processing area for precise operations [3]. Deep learning is based on multilevel learning of various representations on a hierarchical structure. In deep learning, the top level concepts are defined based on low level features and the low level concepts help to define top level concepts [3]. The main purpose of deep learning is to extract intelligently the features in several learning steps.

There are different types of DNNs such as convolutional neural network (CNN), deep belief neural network (DBNN), long short-term memory and etc. One of the DNNs used in the image processing is the CNN which is employed in this study. In the following, we will describe it.

3.1. Convolutional neural networks

The inputs of CNNs are images. In the structure of this type of networks, the main methods such as weight sharing, local receptor field and spatial down sampling is used. This network is partly similar to a biological vision system. Conceptually, the CNNs can be considered as a system of multiple hierarchical detectors connected together.

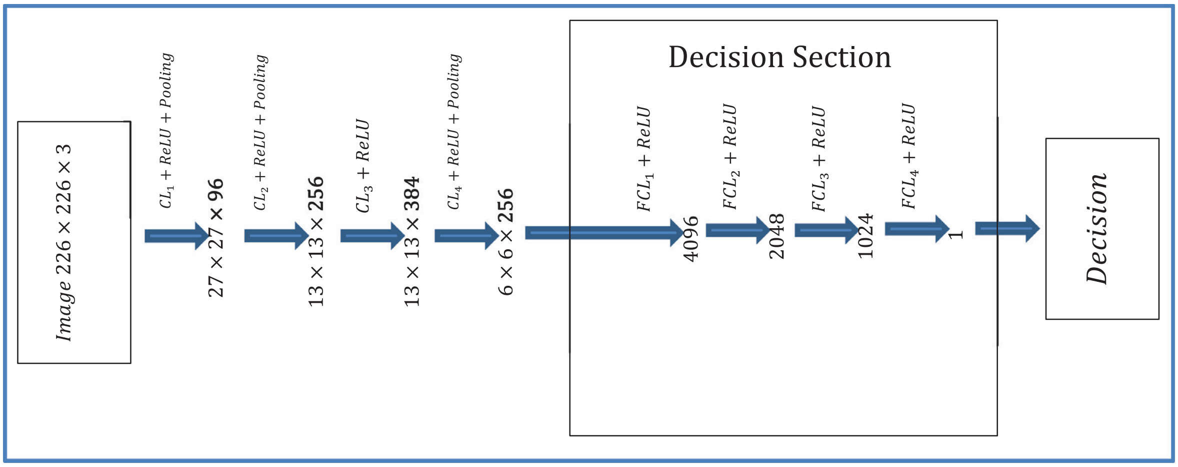


Fig. 2. The proposed CNN architecture for discriminating unconventional images from conventional images.

In the first layer of this network, regularly non-linear template matching is done with proper location resolution and the main features of the input data are extracted. Since a kernel function with a neighborhood locality for each point is used, the template matching is nonlinear. Next layer recognizes the presence of a spatial combination of previous features. Down sampling is sometimes accompanied by this CNN. In this case, next layer performs pattern recognition on a wider spatial scale and a lower resolution. The large number of inputs can be processed by using the down sampling characteristic of the CNNs in spite of their limited amount of tunable weights.

The main core of the CNNs is the convolutional layer. This layer consists of neurons in the form of three dimensional masses with specific width, elevation and depth. The number of neurons per mass is determined depending on the input image space. The output of this layer forms a mass with spatial distribution that is processed by the next convolutional layer and it is finally transferred to the output layer. In fact, the next layer is obtained by sliding a filter on the current layer mass.

The existence of down sampling or pooling layer among other layers decreases the clarity of the subsequent layers. The pooling layer is usually located between convolutional layers. This pooling layer reduces image size (and subsequently computational cost of DNN) and it also controls overfitting. The pooling layer operates independently on every depth slice of the input mass and resizes it in terms of location. On the other hand, it should be mentioned that this layer has the property of spatial translation invariance. In fact, it is reliable to displacement. There are many polling functions such as maximization and averaging functions.

In general, activation function layer is located before the polling layer. In this layer, an activation function is applied on each neuron. As mentioned before activation function is used as a rectified linear unit (ReLU). In general, the normalization layer is located after this layer. A detailed description is provided in Section 3.2. The next layer in these networks is fully connected layer (FC). Neurons in a fully connected layer have full connections to all neurons in the previous layer (as seen in regular ANNs). Their activations can be computed with a matrix multiplication followed by addition of a bias offset.

According to the above explanation, a CNN consists of input, convolution, activation function, normalization, pooling and fully connected layers. In this type of network, the number of convolutional layers and pooling layers, the size of filters, the number of neurons and the connect structure between neurons, should be exactly specified. Common back propagation used in training regular ANNs can also be employed for training of a CNN [11].

LeNet consists of a convolutional layer followed by a pooling layer, another convolutional layer followed by a pooling layer and then three fully connected layers. AlexNet network contains a convolutional layer followed by a pooling layer, 2 convolutional layers, 2 pooling layers and 2 fully connected layers respectively. In fact, in the outdated networks, it is common to locate the pooling after convolutional layer. In this study, proposed method is based on AlexNet architecture and it is presented in the following.

3.2. Proposed CNN architecture

Design of the network architecture is very influential in the performance of proposed system. In the proposed method, a new architecture is presented by the combination of AlexNet and LeNet architecture. In this new method, the convolutional and pooling and fully connected layers are used. Parts of selection of each layer are like AlexNet where several convolutional layers respectively with activation function ReLU have been used. In this architecture, some parts are like LeNet where each convolutional layer is followed by a pooling layer. It has been empirically shown that this architecture has a good performance in unconventional images recognition. In the following, we explain the structure of the proposed architecture depicted also in Fig. 2. The proposed architecture is as follows. The proposed architecture contains several parts. Each of the first two successive parts includes a convolutional layer, a ReLU activation function and a pooling task. The third part contains a convolutional layer and a ReLU activation function layer. The fourth part includes a convolutional layer, a ReLU activation function and a pooling task. Each of the last three parts includes a fully connected layer and a ReLU activation function.

The activation function layer applies an activation function to all neurons. This function doesn't change the size of mass in previous step. ReLU activation function with input of x is shown as $max(0, x)$. ReLU activation function improves the convergence rate of random descending gradient than sigmoid function and hyperbolic tangent. This approach is simpler than other functions [11].

In the proposed network, the input image size is of size 226 × 226 × 3 and each of the 96 used filters is of size 10 × 10 × 3 along with stride 4. Therefore, the output of first layer is a mass of the size 55 × 55 × 96 (note that $55 = \frac{226-10}{4} + 1$). The number of parameters in this layer is 28,896 (note that 28,896=(10 × 10 × 3 + 1) × 96). The number of computations is 87,410,400 in this layer (note that 87,410,400=28,896 × 55 × 55). Then, ReLU activation function is applied. After that, at the maximum pooling task in the first layer a filter 3 × 3 with the stride 2 is applied. Therefore, the output mas of the first layer after maximum pooling task is of the size 27 × 27 × 96 (note that $27 = \frac{55-3}{2} + 1$). Then, a normalization is done.

The input mass size of the next layer is 27 × 27 × 96 and each of the 256 used filters is of size 5 × 5 × 96 along with stride 1 and padding 2. Therefore, the convolutional output of the second layer is a mass of the size 27 × 27 × 256. The number of parameters in this layer is 614,656 (note that 614,656=(5 × 5 × 96+1) × 256). The number of computations is 448,084,224 in this layer (note that 448,084,224=614,656 × 27 × 27). Then, after ReLU activation function, the maximum pooling task in the second layer is just like the maximum pooling task in the previous layer; therefore, the output mas of the second layer after maximum pooling task is of the size 13 × 13 × 256 (note that $13 = \frac{27-3}{2} + 1$). Then, another normalization is done.

The input mass size of the next layer is 13 × 13 × 256 and each of the 384 used filters is of size 3 × 3 × 256 along with stride 1 and padding 1. Therefore, the convolutional output of this layer is a mass of the size 13 × 13 × 384. The number of parameters in this layer is 885,120 (note that 885,120=(3 × 3 × 256+1) × 384). The number of computations is 149,585,280 in this layer (note that 149,585,280=885,120 × 13 × 13). The maximum pooling task is not available in this layer and only ReLU activation function is applied; therefore, the output mas of this layer is of the size 13 × 13 × 384.

The input mass size of the next layer is 13 × 13 × 384 and each of the 256 used filters is of size 3 × 3 × 384 along with stride 1 and padding 1. Therefore, the convolutional output of this layer is a mass of the size 13 × 13 × 256. The number of parameters in this layer is 884,992 (note that 884,992=(3 × 3 × 384+1) × 256). The number of computations is 149,563,648 in this layer (note that 149,563,648=884,992 × 13 × 13). ReLU activation function is applied and then the maximum pooling task just like the first two layer is applied; therefore, the output mas of this layer is of the size 6 × 6 × 256 (note that $6 = \frac{13-3}{2} + 1$). The next layers like a perceptron neural network have fully connected layers. Each layer contains 4096, 2048 and 1024 neurons. It means that the first fully connected layer contains 4096 neurons. The number of parameters and the number of computations are $37, 748, 736$ (note that $37, 748, 736 = 6 \times 6 \times 256 \times 4096$). Then, ReLU activation function is applied. The second fully connected layer contains 2048 neurons. The number of parameters and the number of computations are $8, 388, 608$ (note that $8, 388, 608 = 4096 \times 2048$). Then, ReLU activation function is applied again. The third fully connected layer contains 1024 neurons. The number of parameters and the number of computations are $2, 097, 152$ (note that $2, 097, 152 = 2048 \times 1024$). Then, ReLU activation function is applied again.

The result of all convolutional operations is equal to a large matrix multiplication. This operation produces the result of multiplication point between all filters and points of the receptive fields. All parameters of convolutional layers and those of fully connected layers are respectively 2,413,664 and 48,234,469, i.e. 4.77% and 95.23% of total parameters. All forward computations of convolutional layers and those of fully connected layers are respectively 834,643,552 and 48,234,469, i.e. 94.54% and 5.46% of total computations.

Finally, the output layer includes the output and rating categories. In this paper, the output layer is the rating of conventional and unconventional categories. In the application of intelligent filtering of images, the images can be categorized as *porn, non-porn* and *porn-like* images. According to our study, the images are categorized as conventional and unconventional images. The *porn* and *porn-like* images are assumed to be in the unconventional images category in this article.

In general, the normalization of data is performed on the output of the activation function [11]. The ReLU activation function have the desirable property that they do not require input normalization to prevent them from saturating. However, the normalization presented in Eq. (1) is used to more generalizability of this algorithm.

$$\dot{v}^i_{xy} = \frac{v^i_{xy}}{\left(C + a\sum_{j=\max\left(0,i-\frac{n}{2}\right)}^{\min\left(K-1,i+\frac{n}{2}\right)} \left(v^j_{xy}\right)^2\right)^B} \tag{1}$$

where v^i_{xy} is the activity of a neuron computed by applying ith kernel at position (x, y) and $\dot{v}^i_{xy}$ is the normalized v^i_{xy}, the parameter K is the total number of kernels in the current layer, the constants C, n, a and B are hyper parameters whose values are determined using a validation set. In this paper, we used $C = 2$, $n = 5$, $a = 10^{-4}$, $B = \frac{3}{4}$. The pooling layer decreases spatially the input image in each deep cutting. As it mentioned before, there are many pooling function for convolutional networks which the most introduced of them is the maximum (or max) polling function. This function is a non-linear down-sampling function that is used to identify the most important features. Batch size is an important factor affecting performance of deep learning based methods and should be selected carefully [14]. Training algorithm in DNN is also important. DNN is trained by stochastic gradient descent, in this paper [14]. Batch size is also set as noted by [15].

Three types of the proposed model have been used: (1) the proposed model with randomly initialized weights and fully trainable weights denoted by random fully trainable proposed model (RFTPM), (2) the proposed model the weights of whose four CLs are initialized by AlexNet and with fully trainable weights [11] denoted by fully trainable proposed model (FTPM), and (3) the proposed model the weights of whose four CLs are initialized by AlexNet [11] and frozen throughout training phase denoted by half trainable proposed model (HTPM).

4. Experimental study

In this part, data collection, the evaluation metrics, the baseline methods and comparison of the proposed methods and other methods in unconventional images recognition are presented. The experimentations of this article have been performed on a computer with core i7 2.67 GHz, 4 cores, and 4 GBs of memory.

4.1. Benchmark datasets

The NPDI porno dataset [16] consists of about eighty hours of 400 porno videos and 400 non-porno ones. 10,340 conventional images and 6387 unconventional ones have been extracted totally from these videos. We name this dataset as NPDI.

118,305 porn (unconventional) images and 302,177 non-porn (conventional) ones have been gathered into a handmade dataset. The unconventional images have been collected from some porno websites through web-crawler. After removing conventional images among them by hand, 98,361 unconventional images remain. The conventional images also contain all types of images: those with partially undressed people (as hard samples), those with normally dressed people (as easy samples), and those without people (as very easy samples). A 98,361 randomly chosen conventional images are added to 98,361 unconventional images. We consider it as our handmade dataset.

Each dataset is randomly partitioned into three parts with the sizes of 52.8%, 13.2% and 34%. The part with 52.8% of all samples is considered as training set. The part with 13.2% of all samples is considered as validation set. The last part is used as test set.

4.2. Evaluation criteria

The quantitative criteria such as precision rate (PR), recall rate (RR), accuracy rate (AR) and f-score rate (FR) are used in order to demonstrate the recognition performance of the proposed method in unconventional images. These criteria are used in several studies for evaluation of unconventional images recognition methods [10]. PR criterion is based on the prediction accuracy of the classifier. It also indicates the amount of classifier output confidence. It is computed based on Eq. (2).

$$PR = \frac{TP}{TP + FP} \tag{2}$$

where TP (true positive) is the number of the unconventional images which are assigned to unconventional class by classifier and FP (false positive) is the number of the conventional images which are wrongly assigned to unconventional class. RR criterion indicates the performance of the classifier according to the occurrence of a particular class. It is computed based on Eq. (3).

$$RR = \frac{TP}{TP + FN} \tag{3}$$

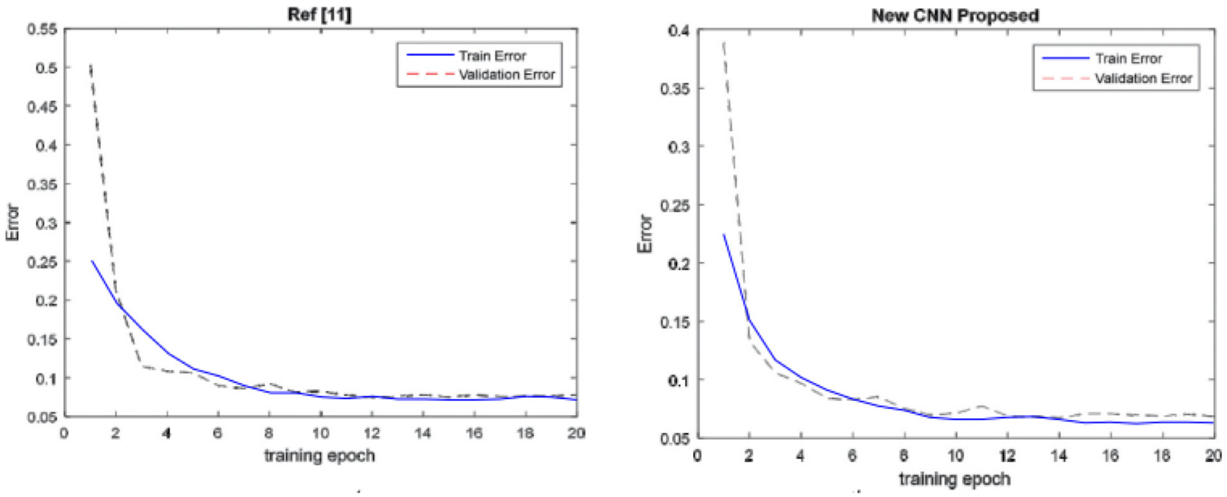
where TP is true positive and FN (false negative) is the number of the unconventional images which are wrongly assigned to conventional class. FR criterion is the combination of PR and RR criteria. The FR criterion is computed based on Eq. (4).

$$FR = \frac{2 \ \times RR \times PR}{RR + PR} \tag{4}$$

The AR criterion is computed based on Eq. (5).

$$AR = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

A receiver operating characteristic (ROC) curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. It is a comparison of two operating characteristics (TP rate and FP rate) as the criterion changes. The ROC curve is also another method of our evaluation. The ROC area under curve (AUC) is also another used metric for our experimentations.



(a) (b)

Fig. 3. The training and validation errors during learning on (a) AlexNet (b) the CNN RFTPM.

Table 1
Comparing the confusion matrices of the proposed CNN (i.e. RFTPM) and AlexNet on the test data of NPDI (top) and handmade (down) da Table 2 tasets.

	Predicted as unconventional		Predicted as conventional	
	FTPM	AlexNet	FTPM	AlexNet
Real unconventional	TP		FN	
	2079	2030	93	142
Real conventional	FP		TN	
	144	170	3372	3346
Real unconventional	TP		FN	
	32,550	32,001	893	1442
Real conventional	FP		TN	
	786	1236	32,657	32,207

4.3. Baseline methods

Bag-of-Visual-Words (BoVW) along with principal component analysis (PCA) on SIFT features and Gaussian Mixture Model (GMM) [17] denoted by BoVW1, BoVW along with support vector machine (SVM) with polynomial kernel [18] denoted by BoVW2, Mask-SIFT with classifier ensemble [19] denoted by Mask-SIFT, Random Forest [1] denoted by RF, ANN [20] denoted by ANN, BoVW along with ORB features [21] denoted by BoVW3, CNN with CaffeNet architecture [22] denoted by CaffeNet, CNN with ImageNet architecture [11] denoted by ImageNet, SIFT features based classification [12] denoted by SIFT, texture and visual features based model [23] denoted by TVFM, skin-shaped features based model along with classifier ensemble [24] denoted by SSFM, and content based image recovery model [25] denoted by CIRM are the baseline of this work.

4.4. Experimental results

In the used benchmark dataset, two third of all data points are considered as training data during training phase and one third of them are considered as test data during testing phase. In the training phase, one third of training data points are used for validation task.

At first, there is a comparison between famous architecture AlexNet and the proposed architecture. Fig. 3 indicates the obtained results of classification in the training and validation stages with the proposed CNN architecture and AlexNet [11]. As shown in Fig. 3, the proposed CNN (i.e. RFTPM) outperforms AlexNet in the training data. The proposed architecture has shown less error rate than the AlexNet method. The comparison of both architectures is shown in Table 1 separately on the test data on NPDI benchmark. TP rate and FP rate criteria are used for comparison. TP rate is the ratio of correctly identified unconventional images to the total unconventional images. FP rate is the ratio of wrongly identified conventional images to the total conventional images. Table 1 separately shows the results of the same models on the test data on handmade benchmark.

Table 2
The experimental results of different scenarios of the proposed CNN comparing with state of the art methods on NPDI dataset.

Measure	FR	AR	RR	PR	Time per image in second	
Method					Train time	Test time
TLANet	89.20%	89.27%	88.62%	89.78%	0.07	0.090
AlexNet	92.86%	92.82%	93.46%	92.27%	0.63	0.089
HTPM	92.86%	92.80%	93.63%	92.11%	0.06	0.083
RFTPM	94.61%	94.54%	95.72%	93.52%	0.56	0.075
FTPM	95.06%	95.00%	96.19%	93.96%	0.36	0.076

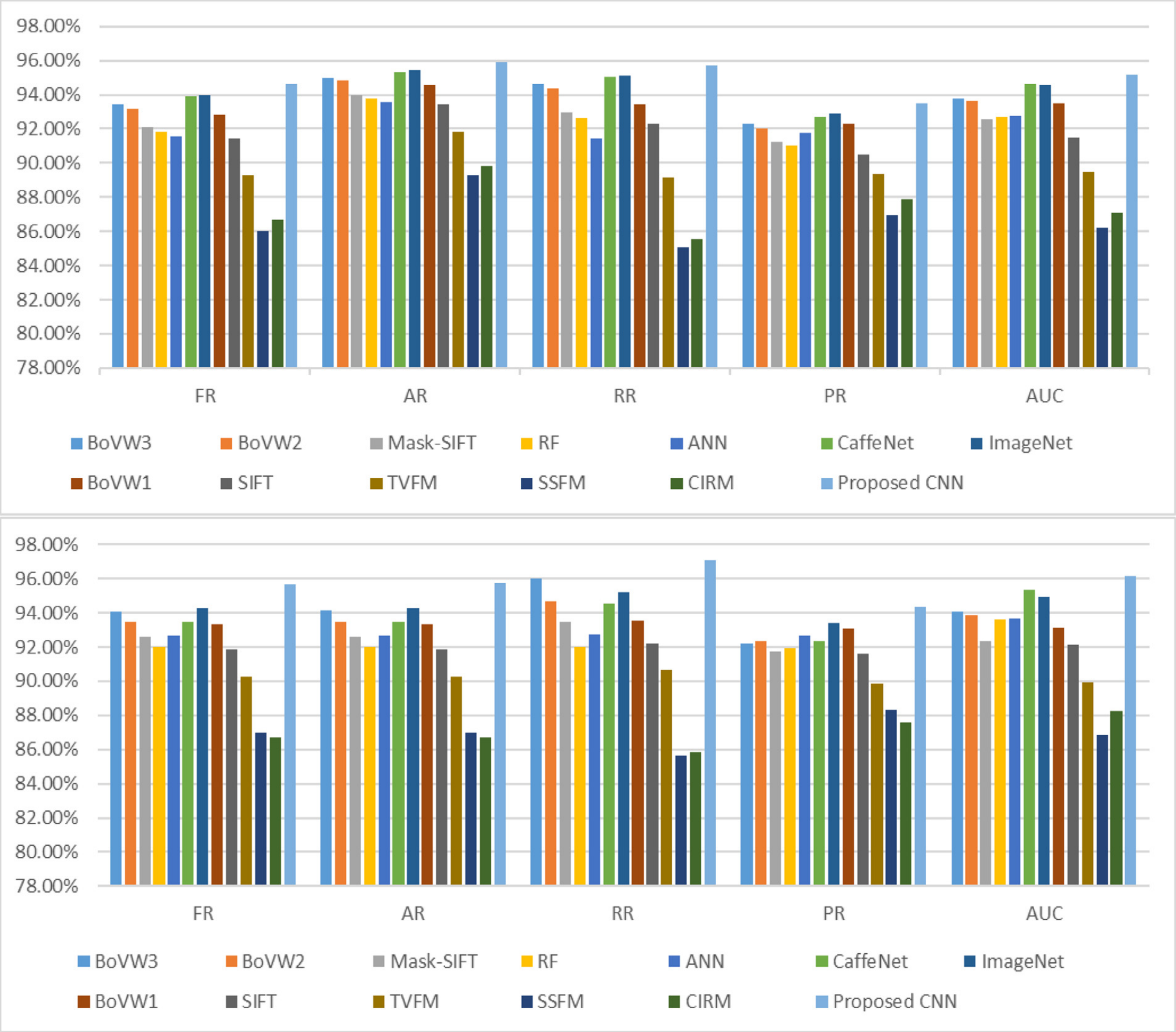


Fig. 4. The results of the proposed CNN (i.e. HTPM) against other state of the art methods in terms of different criteria on the test data of NPDI (top) and handmade (down) datasets.

According to Table 1, the obtained result of the proposed CNN architecture (i.e. RFTPM) has a better performance than AlexNet architecture on the testing data. Table 2 shows the criteria introduced in the evaluation section in all proposed architectures and traditional AlexNet and transfer learning AlexNet (TLANet). The CNN with AlexNet and TLANet architectures are used as the main configuration [11], because, there has not been any research for recognition of unconventional images through a CNN with AlexNet architecture. TLANet indicates the AlexNet where only the last layer is trainable and other layers are initialized by its traditional weights from [11]. The results indicate that convolutional networks with the proposed architecture have better performance than AlexNet and TLANet architectures. Among the RFTPM, FTPM, and HTPM, the FTPM

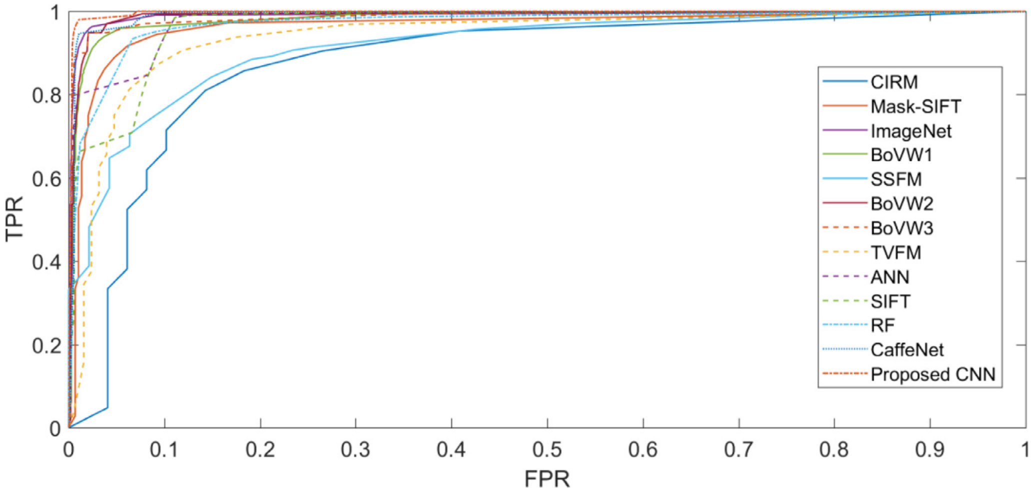


Fig. 5. The ROC of the proposed CNN (i.e. HTPM) against other state of the art methods on handmade benchmark dataset.

is superior to others. Table 2 presents the consumed time in training and test phases for each image. The comparison of running time of algorithms indicates that performance of the proposed CNNs is improved in comparison to the reference methods at the training and test phases. Indeed, FTPM has an acceptable consumed time in training and test phases for each image comparing to AlexNet and RFTPM architectures. While HTPM and TLANet have less consumed time in training and test phases for each image, they are weaker in performance. From here on, HTPM is considered to be the proposed method.

Fig. 4 indicates the comparison of the proposed CNN method and the other state of the art methods for recognition unconventional images. According to the obtained results, the DL method has suitable performance, and the CNN with the proposed architecture has shown the best performance. In the intelligent filtering of images in addition to correctly identifying the unconventional images, it is so important to correctly recognize conventional ones. Therefore, the proposed method performs better than other presented methods.

AR is not sufficient to confirm which method is superior, because there are many hidden factors during experimentations. A nonparametric Wilcoxon test is used to demonstrate if the proposed method is superior to others with a *p*-value lower than 0.05 (i.e. 95.00% confidence test). Our experiment is repeated 30 times as our samples during the test. The p-values for the accuracy of different models on NPDI and handmade datasets are respectively 0.04 and 0.03. ROC curve of different models and proposed CNN have been depicted in Fig. 5 on handmade dataset.

5. Conclusions and future work

Vast increase of internet access and significant growth of web based broadcasters have currently resulted in distribution and sharing of informative resources such as worldwide images. Although this kind of sharing may bring many advantages, there are also disadvantages, such as access of kids to porn images, which should not be neglected. While models based on deep learning have been among the best ones in image and vision understanding tasks, its application in pornographic filtering has been ignored. Therefore, in this paper, an attempt is made to propose an approach for classification of the unconventional images and their intelligent filtering employing a deep learning based model. In this study, a method based on convolutional neural network with a new architecture is proposed. A standard NPDI dataset and a handmade dataset have been used as benchmark. The obtained results on two benchmarks indicate that the proposed method is far more accurate than the other methods recently developed in terms of many criteria. Also, by an ablation analysis, we investigate the effectiveness of each subsection in our method in detail. The statistical nonparametric Wilcoxon test is used and it approves the superiority of our method to the other ones is significant. We showed that the proposed method is not only the superior one in terms of accuracy among all compared method, it is also more the superior one in terms of scalability among them. As a future work, effect of using an ensemble of our model can be investigated. Using other deep models can be another direction for future work. Using simultaneously deep features and handmade features can be considered to be another direction for future work too.

Declaration of Competing Interest

The authors declare that they have no financial or non-financial competing interests.

CRediT authorship contribution statement

Huan Niu: Formal analysis, Resources, Software, Visualization, Validation, Writing - original draft, Writing - review & editing. **Wei Xu:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Software, Supervision, Writing - original draft, Writing - review & editing. **Hamidreza Akbarzadeh:** Formal analysis, Resources, Software, Writing - original draft. **Hamid Parvin:** Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Resources, Software, Validation, Supervision, Writing - original draft, Writing - review & editing. **Amin Beheshti:** Validation. **Hamid Alinejad-Rokny:** Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Supervision.

Ethics approval and consent to participate

Not applicable.

Consent for publication

All authors have read and approved the final version of the paper.

Availability of data and material

All data are publicly available.

Funding

This work was supported by grants to HAR and HP. HAR is currently supported by the UNSW Scientia Fellowship.

References

[1] Zuo H, Hu W, Wu O. Patch-based skin color detection and its application to pornography image filtering. In: Proceedings of the 19th international conference on World wide web. ACM; 2010.

[2] T. Largillier, G. Peyronnet, and S. Peyronnet, Efficient filtering of adult content using textual information. Murdock et al. [7], arXiv:1512.00198, 2016: p. 14–17.

[3] Goceri E. Formulas behind deep learning success. In: Int. Conf. on Applied Analysis and Mathematical Modeling (ICAAMM2018); 2018. p. 156. pgJune 20-24.

[4] Niu H, Khozouie N, Parvin H, Alinejad-Rokny H, Beheshti A, Mahmoudi M. An Ensemble of Locally Reliable Cluster Solutions. Appl. Sci. 2020;10(5):1891. doi:10.3390/app10051891.

[5] Chang Z, Cao J, Zhang Y. A novel image segmentation approach for wood plate surface defect classification through convex optimization. J Forestry Res 2018;29(6):1789–95.

[6] Fleck MM, Forsyth DA, Bregler C. Finding naked people. European Conference on Computer Vision. Springer; 1996.

[7] Zhuo L, Zhang J, Zhao Y, Zhao S. Compressed domain based pornographic image recognition using multi-cost sensitive decision trees. Signal Processing 2013;93(8):2126–39.

[8] Li FF, Luo SW, Liu XY, Zou BJ. Bag-of-visual-words model for artificial pornographic images recognition. J Central South Univ 2016;23(6):1383–9.

[9] Yin H, Huang X, Wei Y. SVM-based pornographic images detection. In: Software Engineering and Knowledge Engineering: Theory and Practice. Berlin, Heidelberg: Springer; 2012. p. 751–9.

[10] Ries CX, Lienhart R. A survey on visual adult image recognition. Multimed Tools Appl 2014;69(3):661–88.

[11] A. Krizhevsky, I. Sutskever, and G.E. Hinton. Imagenet classification with deep convolutional neural networks. in Advances in Neural Information Processing Systems. 2012.

[12] S.M. Kia, H. Rahmani, R. Mortezaei, M.E. Moghaddam, and A. Namazi, A novel scheme for intelligent recognition of pornographic images. arXiv preprint arXiv:1402.5792, 2014.

[13] Islam M, Watters P, Yearwood J, Hussain M, Swarna LA. Illicit image detection using erotic pose estimation based on kinematic constraints. In: Innovations and Advances in Computer, Information, Systems Sciences, and Engineering. Springer; 2013. p. 481–95.

[14] Goceri E, Gooya A. Int. Conf. on Mathematics (ICOMATH2018), An Istanbul Meeting for World Mathematicians, Minisymposium on Approximation Theory & Minisymposium on Math Education Istanbul, Turkey; 3-6 July 2018.

[15] Javanmard R, JeddiSaravi K, Alinejad-Rokny H. H. proposed a new method for rules extraction using artificial neural network and artificial immune system in cancer diagnosis. J Bionanosci 2013;7(6):665–72.

[16] Malamuth NM. Criminal and noncriminal sexual aggressors. Ann NY Acad Sci 2003;989(1):33–58.

[17] Deselaers T, Pimenidis L, Ney H. Bag-of-visual-words models for adult image classification and filtering. In: International Conference on Pattern Recognition (ICPR); 2008. p. 1–4.

[18] Ulges A, Stahl A. Automatic detection of child pornography using color visual words. In: 2011 IEEE International Conference on Multimedia and Expo; 2011. p. 1–6.

[19] Steel CM. The Mask-SIFTcascading classifier for pornography detection. In: World Congress on Internet Security (WorldCIS); 2012. p. 139–42.

[20] Zaidan AA, Ahmad NN, Larbani HAM, Zaidan BB, Sali A. On the multi-agent learning neural and bayesian methods in skin detector and pornography classifier: an automated anti-pornography system. Neurocomputing 2014;131:397–418.

[21] Zhuo L, Geng Z, Zhang J, Guang Li X. ORB feature based web pornographic image recognition. Neurocomputing 2016;173:511–17.

[22] Nian F, Li T, Wang Y, Xu M, Wu J. Pornographic image detection utilizing deep convolutional neural networks. Neurocomputing 2016;120:283–93.

[23] Parvin H, MirnabiBaboli B, Alinejad-Rokny H. Proposing a classifier ensemble framework based on classifier selection and decision tree. Eng Appl Artif Intell 2015;37:34–42.

[24] Zheng QF, Zeng W, Wang WQ, Gao W. Shape-based adult image detection. Int J Image Graph 2006;6(1):115–24.

[25] Shih JL, Lee CH, Yang CS. An adult image identification system employing image retrieval technique. Pattern Recognit Lett 2007;28(6):2367–74.

Ms Huan Niu was born in Anhui, China, in 1989. Currently, she is working in acoustics research group of Communication University of China to pursue her doctor's degree in signal and information processing. Her primary research is focused on research on audio and video algorithm. Email: niuhuan@cuc.edu.cn.

Mr. Wei Xu is a student at Hubei collaborative innovation center of digital forensics and trusted application, China. His-research interest is machine learning and deep learning.

Mr. Hamidreza Akbarzadeh is a student at department of computer science, Yasooj branch, Islamic Azad University, Yasooj, Iran. His-research interest is machine learning, data mining and deep learning.

Dr. Hamid Parvin is an assistant professor at department of computer science, Nourabad Mamasani branch, Islamic Azad University, Fars, Mamasani, Iran. His-research interest is machine learning, data mining and deep learning. dr. Parvin has published more than 50 research papers in machine learning and data mining.

Dr. Amin Beheshti is the Director of AI-enabled Processes (AIP) Research center and the head of Data Analytics Research Lab, Macquarie University. In addition to his contribution to teaching activities, Amin extensively contributed to research projects; where he was the R&D Team Lead and Key Researcher in the 'Case Walls & Data Curation Foundry' and 'Big Data for Intelligence' projects.

Dr. Hamid A. Rokny is a member of UNSW Graduate School of Biomedical Engineering (GSBME). He is also Heath Data Analytics Program Leader of AI-enabled Processes (AIP) Research center. Dr. Rokny's Lab at GSBME focuses on using cutting-edge machine learning and systems biology techniques in conjunction with health and medical data to better understand and interpret medical and health patterns.